TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Kevin Kanarbik

Farmakogeneetilised variandid täisgenoomsetes andmetes

Bakalaureusetöö (9 EAP)

Juhendaja: Tõnis Tasa, MSc

TARTU 2017

Farmakogeneetilised variandid täisgenoomsetes andmetes

Lühikokkuvõte: DNA sekveneerimine ehk orgaaniliste molekulide järjestamine on muutumas üha lihtsamaks ja odavamaks. See lihtsustab laiahaardeliste sekveneerimisprojektide läbiviimist, mille üheks rakenduseks on erinevate geograafiliste piirkondade inimeste geneetika võrdlemine. Farmakogeneetilisteks variantideks nimetatakse asukohti genoomis, mida on senistes teadusuuringutes suudetud seostada mõningase mõjuga ravimitoimetele. Kõigi geneetiliste, sealhulgas farmakogeneetiliste variantide eri populatsioonide sageduste vahel esineb variatsioone. Käesoleva bakalaureusetöö eesmärk on kirjeldada Eesti Geenivaramust saadud eestlaste valimis esinevaid farmakogeneetilisi variante. Antud variantide sagedusandmeid kasutatakse, et võrrelda erinevate populatsioonide omavahelist sarnasust ja erisust.

Võtmesõnad: farmakogeneetika, populatsioonigenoomika, andmeanalüüs

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Pharmacogenetical variants in whole-genome data

Abstract: Human DNA sequencing is becoming easier and cheaper. This makes it simpler to conduct a wide range of sequencing projects. These can help compare genetic data from different geographical populations. Pharmacogenetic variants are genetic elements that scientific studies have linked to effects on drug metabolisation. All genetic variants including pharmacogenetic variants show some between-population frequencies. The main aim of this bachelor's thesis is to describe pharmacogenetic variants that appear in the Estonian population and to use pharmacogenetic variant frequencies from whole-genome sequencing data of Estonian Genome Center in order to describe similarities and differences in various populations.

Keywords: Pharmacogenetics, population genomics, data analysis

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

Sisukord

1. Sissejuhatus

Inimeste geneetiliste andmete suurimine on muutumas järjest lihtsamaks ja odavamaks [1]. Nüüd on võimalik võrrelda erinevate piirkondade inimeste DNA järjestusi. Inimolendid on omavahel 99% ulatuses geneetiliselt identsed, kuid ülejäänud 1% tagab inimeste unikaalsuse. [2]. Tänased valdkondsed uuringud keskenduvad põhiosas vaadeldud geneetiliste variatsioonide seostamisega inimestele omaste fenotüüpide ja ilmutatud tunnustega. Inimestevaheline bioloogiline erinevus tagatakse selle varieeruvuse poolt. Farmakogeneetika on teadusharu, mis tegeleb geneetiliste variandide mõju uurimisega ravimite metabolismile. Teaduslikes uuringutes ravimi metabolismiga seostatud variante nimetatakse *farmakogeneetilisteks variantideks*. Teatud ravimite mõju nende kandjatele erineb mittekandjatele avalduvast mõjust. Farmakogeneetika üheks eesmärgiks on ravimite väljakirjutamist personaliseerida, nii et patsientidel oleks võimalik tarvitada vaid neile sobivaid ravimeid [3].

Antud bakalaureusetöö eesmärk on kirjeldada eestlaste seas levinud farmakogeneetilisi variante ja võrrelda neid valitud hulga teistest populatsioonidest pärit sarnaste andmetega. Tuvastame eestlaste seas levimuselt teiste populatsioonidega võrreldes ekstreemseid variante ja anname kirjanduse põhjal ülevaate nende teadaolevast mõjust. Töö käigus ühendatakse erinevatest allikatest pärit andmebaase ja võrreldakse erinevate rahvaste geneetiliste variantide sagedusi. Võrdlemisel rakendatakse erinevaid statistilisi ja matemaatilisi meetodeid. Töös kirjeldatakse eestlaste ning teiste võrdlusaluste populatsioonide sarnasusi ja erinevusi. Peamiseks hüpoteesiks on, et geograafiliselt lähedasemad populatsioonid on farmakogeneetiliste variantide profiililt üksteisele sarnasemad.

Käesoleva töö eeskujuks on analoogilise ülesehitusega uurimustööd. Inimgeneetika instituudis Pittsburghi Ülikoolis avastati, et ülekaaluliste arvuga silmapaistvatel samoalastel on teiste rahvastega võrreldes rohkem levinud rs373863828 geneetiline variant, mis asub geenis *CREBRF*. See variant on seotud ülekaalulisuse tekkimisega ja on teiste populatsioonide seas haruldane, kuid esineb 25% samoalastel [4]. Samuti avastati Mehhikos Rahvuslikus Meditsiini ja Toitumise Instituudi uuringus, et mehhiklastel on ka ülekaalulisuse probleem, mis on mingil moel tingitud geenidest. Nimelt avastati, et esineb seos kehamassiindeksi ja *11q13* geenis

asuva geneetilise variandi rs614080 vahel [5]. On näidatud, et teatud BRCA1 või BRCA2 geenialleele kandvad naised haigestuvad 60% suurema tõenäosusega rinnavähki [6] . Maroko tuberkuloosi patsiente uurides on samuti selgunud, et inimestel on tuberkuloosi nakatumise tõenäosust tõstab *CYP7A1* geenis asuv variant rs3808607. [7]

2. Teoreetiline taust

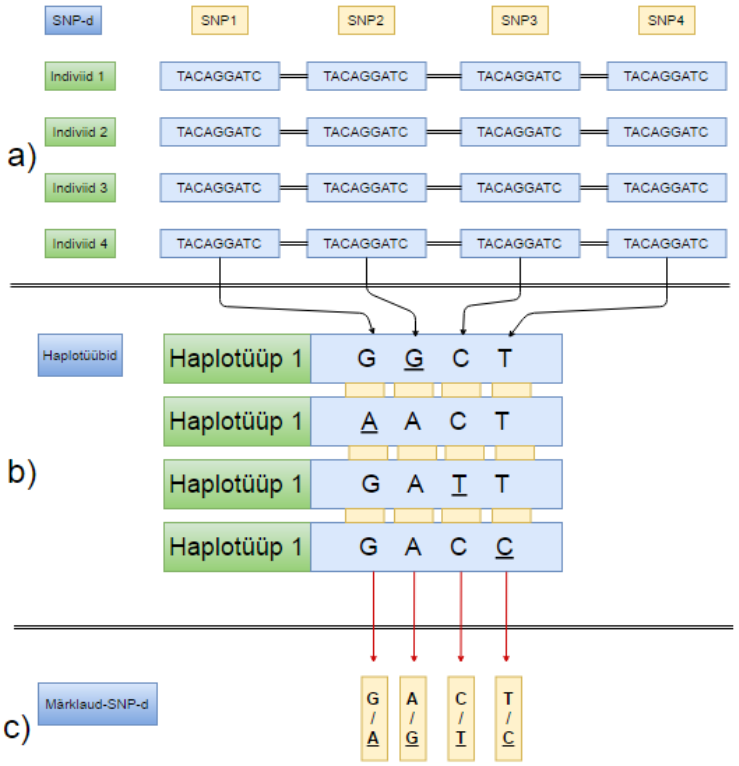
Esimese peatüki eesmärk on lugejale tutvustada töö geneetilist ja statistilist tausta. Seletatakse lahti töös kasutatavad mõisted ja terminid.

2.1. Geneetika

Geneetika on teadusharu, mis uurib organismide pärilikkust, selle variatsioone ja muutlikkust. Inimgeneetika on õpe variatsioonidest, mis esineb inimestel. Elavad organismid koosnevad rakkudest, mis sisaldavad geneetilist materjali ehk pärilikkusainet. Seda säilitab rakk enda tuumas ja kannab edasi võimalikult muutumatult järglasrakkudesse. Rakkude tuumas asuvad kromosoomid, mis on pärilikkuse informatsiooni kandjad. Inimeste rakud koosnevad kahest 23 kromosoomist koosnevast komplektist, kus üks pool on saadud emalt ja teine isalt. Kõik kromosoomid organismis moodustavad kromosoomikomplekti, milles on erinevad geneetilise informatsiooni üksused ehk geenid, mis määravad organismile või rakule mingi tunnuse, näiteks silmade värvus. Organismi terve kromosoomikomplekt on genoom, mis on organismi geneetilise informatsiooni terviklik koopia.

Geneetiline informatsioon asub kromosoomides, mis omakorda koosneb kahest keemilisest makromolekulist (keskmisest suurema massiga molekul): valkudest ja nukleiinhapetest. Viimased on polümeerid, mis koosnevad korduvatest allüksustest ehk nukleotiididest. Neid on omakorda kahte tüüpi: desoksüribonukleiinhape ehk DNA ja ribonukleiinhape ehk RNA. DNA sisaldab nelja põhilist lämmastikualust: adeniin (A), tümiin (T), tsütosiin (C) ja guaniin (G). Desoksüribonukleiinhape esineb inimorganismis tavaliselt kaksikheeliksina, kus kaks ahelat on koos tänu komplementaarsete lämmastikaluste vahelistele vesiniksidemetele, kus adeniin läheb paari alati tümiiniga ja guaniin tsütosiiniga. Hinnanguliselt on inimese genoomi kogupikkus umbes kolm miljardit nukleotiidi- ehk aluspaari.

Alleel on kromosoomi määratud piirkonnas olev üks kahest või mitmest alternatiivsest geeni variandist. Alleeli suhtarv ehk alleelisagedus uuritavas geenis on samas piirkonnas olevate kõikide teiste alleelide suhe populatsioonis. Geenides asuvad mittekodeerivad järjestused ja kodeerivad piirkonnad, mida kutsutakse vastavalt intronideks ja eksoniteks. Kodeerivad DNA piirkonnad transkribeeritakse RNAks, mis omakorda transleeritakse valguks.



Joonis 1. SNPid, haplotüübid ja märklaud-SNPid.

a) Näidatud on nelja inimese kromosoomi samad piirkonnad. Enamik DNA järjestusest on identne, välja on toodud kolm varieeruvat nukleotiidi. Igal SNPil on kaks võimalikku alleeli – esimesel C ja T, teisel ning kolmandal G ja A.

b) Isikute haplotüübid määratud aheldus piirkonnas.

c) On toodud kolm märklaud-SNPsi, mille abil on võimalik määrata geneetilist varieeruvust, ilma et oleks vaja genotüpiseerida haplotüübi kõiki SNPsi. Näiteks, kui konkreetsel kromosoomil oli märgistatud SNPde suhtes järjestus A-T-C, siis see läheb kokku esimese haplotüübiga [8]

Geneetiliste variantide hulka kuuluvad üksiknukleotiidsed polümorfismid, mida nimetatakse SNPideks (ingl *single nucleotide polymorphism aka SNP).* See on DNA järjestuse varieeruvus, mis väljendub ühe nukleotiidi muutumisel genoomis. SNPid ei ole alati organismile kahjulikud. Neid esineb keskmiselt korra 200-300 aluspaari kohta ning mitmetest SNPdest moodustuvat kombinatsiooni ühes kromosoomiosas nimetatakse haplotüübiks. Konkreetse indiviidi haplotüübibloki järjestuse määramiseks piisab vaid mõne iseloomuliku SNP väljaselgitamisest. Selliseid SNPe nimetatakse märklaud-SNPideks. Märklaud-SNPid esindavad teisi samas piirkonnad asuvaid SNPsi, kuna on nendega aheldunud [8].

2.2. Sekveneerimine

Genoomides olevate nukleotiidide järjestuse määramist nimetatakse sekveneerimiseks. DNA sekveneerimisel kasutatakse bioloogilise alusmaterjalina DNAd, mille teostamise tehnoloogiad on ajas pidevalt täiustunud. Esimene põlvkonna Sangeri meetod arendati välja aastal 1975. Sangeri sekveneerimismeetod on jätkuvalt kasutuses teise põlvkonna sekveneerimistulemuste valideerimiseks. Sellega on võimalik järjestada rohkem kui 500 aluspaarilisi DNA lõike. Antud meetod sobib lühikeste DNA ridade nukleotiidide järjestamiseks, kuid selle kasutamine kogu genoomile on tohutult töömahukas. Teise põlvkonna tehnoloogia meetoditega sekveneeritakse tuhandeid molekule paralleelselt. Väljundi pikkus on tavaliselt kuni 150 aluspaari. Teise põlvkonna tehnoloogiat jaotatakse pürosekveneerimiseks ja sünteetiliseks sekveneerimiseks. Teise põlvkonna sekveneerimisel DNA fragmenteeritakse, seejärel valmistatakse ette matriitsid (ingl *template prep*), mille põhjal toimub automatiseeritud sekveneerimine. Kasutusel olevad sekveneerimise tehnoloogiad erinevad üksteisest põhiliselt matriitsi ettevalmistuse poolest.. Saadav väljund vajab põhjalikku kvaliteedikontrolli [9]. Kolmanda põlvkonna tehnoloogiale on omane üksikmolekuli järjestamine reaalajas ilma DNA amplifitseerimiseta[1]. Väljundi pikkus on kuni 10 000 aluspaari [10].

2.3. Täisgenoomid biopankades

Indiviidide DNA järjestuste võrdlusel on võimalik kirjeldada indiviidide lõikes erinevaid variante. Valdav osa seni tehtud geneetilistest uuringutest on keskendunud konkreetsetele geneetilistele piirkondadele. Selleks kasutatakse eriotstarbelisi mikrokiipe, mille põhjal tuvastatakse või genotüpiseeritakse kiibil märgitud individuaalsed variandid. Sekveneerimise tehnoloogiate arenguga on esile kerkinud täisgenoomide sekveerimine, mis on teise põlvkonna sekveneerimise alamliik, mille puhul järjestatakse geneetilist materjali terve inimgenoomi ulatuses. Täisgenoomide sekveneerimise tulemiks on suuremahulised bioinformaatilised järjestusandmed, millele tuleb teostada kvaliteedikontroll. Need tuleb kaardistada genoomile ning tuvastada ja lisada indiviidide vahel erinevatele variantidele annotatsioonid[2]. Varieeruvate positsioonide asukohad on huvipakkuvad selgitamaks mõju väliste tunnuste ehk fenotüüpidega [11].

[1] kordistamine
[2] Märked, mis iseloomustavad tähistust

Täisgenoomide kasutamine on jätkuvalt alternatiividest kallim, kuid on saamas oluliseks tööriistaks haiguslugude uurimiseks geneetiliste vahenditega. Kogutavad andmed võivad aidata avastada uusi patogeenseid variante, mis teiste tehnoloogiatega jääksid avastamata [12]. Üha kasvav hulk uurimisasutusi teostab vastavaid analüüse kohalikus populatsioonis ning lokaalsed biopangad on kohad, kuhu kogutakse populatsioonide bioloogilist materjali, kasutamaks seda uuringutes ja teadustöödes [13]. Eraldiseisvate tööde tulemusi analüüsitakse teadusgruppide üleses koostöös konsortsiumites.

2.3.1. Variant Call Format

Inglise keeles *Variant Call Format aka VCF* on failiformaat, milles hoitakse kaardistatud variante positsioonidega, mille puhul on leitud valimis olevate indiviidide puhul erinevusi. VCF failides sisalduvad sekveneerimisandmete põhjal tuvastatud indiviidide vahel erinevad positsioonid. Failiformaati illustreerib tabel 1. Tabelid on maatriksid, kus igal real on unikaalne geneetiline variant ja veerud sisaldavad variantide infot ning genotüüpe.

Variantide üldine info on kirjeldatud järgnevates veergudes: CHROM tähistab kromosoomi numbrit, POS on geneetilise variandi positsiooni selles kromosoomis, ID näitab, mis on konkreetse geneetilise variandi identifikaatorit ühenukleotiidsete polümorfismide andmebaasis, REF märgib positsiooni referentsalleeli, ALT on alternatiivne alleel, QUAL näitab geneetilise variandi kvaliteediskoori, FILTER tähistab kas geneetilise variant läbis kvaliteedikontrolli filtreid (PASS või FAIL) ja INFO veerus kirjeldatakse variantide lisaannotatsioon nagu alleeli koguarv, sagedus, üldarv jne.

Tabel 1. VCF faili näide

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
14	103539669	rs7149097	A	G	1380620.00	PASS	Multiallelic=0;AC=3716;AN=4488;AF=0.827986;AF…
4	95356328	rs10008257	G	A	1137620.00	PASS	Multiallelic=0;AC=2509;AN=4488;AF=0.559046;AF…
3	12323414	rs2920500	G	A	1080380.00	PASS	Multiallelic=0;AC=2565;AN=4488;AF=0.571524;AF…
4	157011923	rs10030044	G	T	1315160.00	PASS	Multiallelic=0;AC=3020;AN=4488;AF=0.672906;AF…
8	6639024	rs2928608	T	C	113292.00	PASS	Multiallelic=0;AC=331;AN=4488;AF=0.0737522;AF…

Geneetilise variandi rs-koodi (ID) ja alleelisagedust INFO väljalt kasutatakse antud lõputöös. Rs-kood on geneetiliste variantide standardne tähistuskood ning võimaldab variantidega seonduva lisainfo, näiteks variandi seose ravimitega täpsustamist väliste andmebaaside abil. INFO veeru AF tunnus märgistab variandi alleelisagedust ja see on oluline rahvaste vahelise võrdlusel. Erinevate rahvaste variantide alleelisageduse võrdlemisega aitab leida populatsioonide omavahelist sarnasust või erinevust [14].

2.3.2. Eesti Geenivaramu andmed

Eesti Geenivaramu loodi 1999. aastal Eesti Genoomi Projekti Sihtasutuse poolt. Selle projekti eesmärk oli luua bioloogiliste ja meditsiiniliste andmete jaoks biopank uurimaks Eesti populatsiooni geneetikat ning selles levivate haiguste geneetilist ning keskkondlikku komponenti [15].

Uuringus kasutatavad eesti populatsiooni geneetilised sagedused on pärit Eesti Geenivaramu täisgenoomsete andmete sekveneerimise projektist. Täielikult on sekveneeritud 2240 geenidoonorit. Valimis on 977 inimest valitud, kuna nende kohta on kogutud ka muud tüüpi geneetilisi andmeid; 1242 on juhuslikud inimesed kategoriseeritud oma sünnikoha järgi ja geenidoonorite hulgas on ka kaks suurt perekonda ehk 21 inimest [16].

2.3.3. UK10K Project

Suurbritannia ja Põhja-Iiri Ühendkuningriigis (UK) on sekveneeritud peaaegu 10 000 inimese genoomid. Projektiga sooviti iseloomustada variante UK populatsioonis, mis olid haruldased ja seotud varieeruvate geneetiliste haigustega. Kokku leiti üle 42 miljoni ühenukleotiidse variatsiooni. Sekveneeriti 3781 terve indiviidi rakkudest eraldatud genoomi [17].

2.3.4. GoNL Project

GoNL täisgenoomide sekveneerimise projekt oli üks esimesi omataolisi projekte, mille eesmärk oli iseloomustada hollandlaste geneetilist varieeruvust. Viie erineva bioloogilise panga koostöös sekveneeriti 769 hollandlase kogugenoom. Suurema osa valimist moodustasid 231 peret, kus olid esindatud ema, isa ja laps. Lisatud oli ka 19 ühe – või kahemuna kaksikutega peret. Kõik kaasatud indiviidid olid täiskasvanud vanuses 19-87 aastat (keskmine vanus 53) [18].

2.4. Farmakogeneetika

Teadusharu, mis uurib konkreetsete geenide mõju teatud omastamisele, on farmakogeneetika. Geneetiline komponent on oluline kirjeldamaks seoseid erinevate haiguste vahel. Viimase 20 aasta vältel tehtud uuringud on tõestanud, et geneetikal on oluline roll haiguste kujunemisel ja tekkimisel [3]. Farmakogeneetika uurib ravimainete omastamise pärilikkust. See käsitleb, kuidas geneetika mõjutab ravimainete mõju organismile ja ainevahetusele. Farmakogeneetilised variandid on SNPd, mis on seotud mõjuga ravimite metabolismile. Variandikandjatel on muutunud ravimireaktsioon võrreldes mittekandjatega.

Ravimite efektiivsust ja ohutust mõjutavad mitmed geenid. Geneetiliste variatsioonide tausta selgitamine rahvastiku tasemel on geneetikast lähtuva ravi ehk farmakoteraapia oluline uurimisobjekt. Oluline on ravimitele tundlike ja väga resistentsete isikute tuvastamine [19]. Teada on hulgaliselt ravimi või alleeli seoseid, mille puhul on tarvilik ravi määramiseelne geneetiline test otsustamaks individuaalse sobivuse üle. Kliinilise farmakogeneetika rakendamise konsortsiumi suunised (ingl *CPIC aka Clinical Pharmacogenetics Implementation Consortium guidelines*) on autoriteetseimaks kogumiks ravimitarbimiseelset geneetilist testi nõudvatest juhtudest [20]. Ravimite võtmisega on seotud neli protsessi: imendumine, jaotumine, metabolism ja eritumine. Neid nimetakse ADME protsessideks (ingl *Absorption, Distribution, Metabolism and Excretion*). Farmakogeneetikat huvitab ravimi metabolism ehk kuidas ravim kehas lõhustatakse ja laiali kantakse [21].

VIP (ingl *Very Important Pharmacogene)* geenid on alamhulk geene, mille puhul on näidatud suure hulga variantide mõju ravitoimetele. VIP geenide variandid võivad olla näiteks patogeensed (kutsuvad esile haigusi), mittepatogeensed ja ravimitele reageerivad [22].

3. Suuremahulised sekveneerimisprojektid

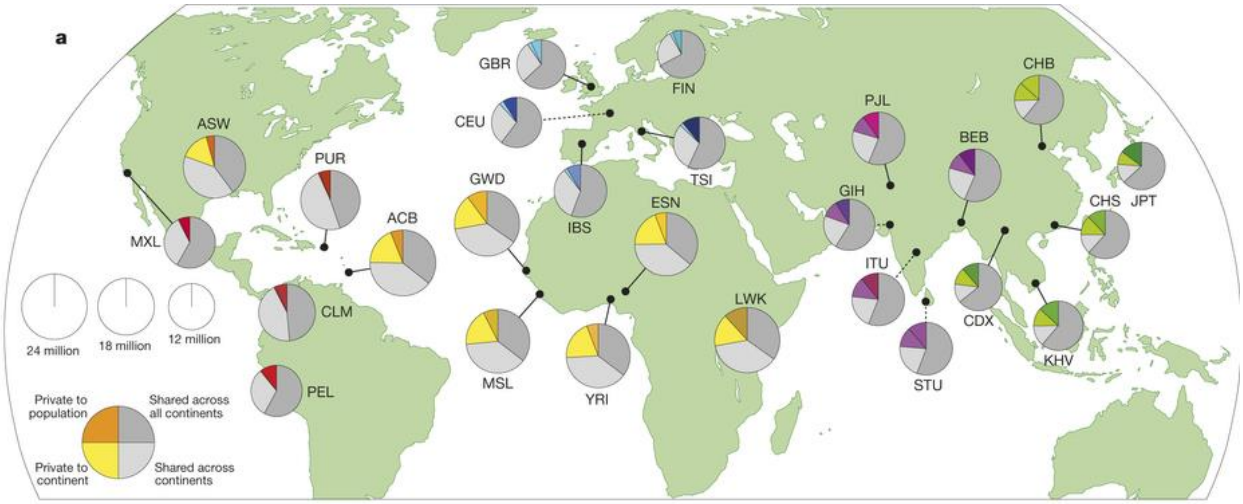
Inimgenoomi projekt oli avalikult rahastatud teadusprojekt, mis kestis 13 aastat. Selle eesmärk oli täielik inimese DNA sekveneerimine. Projekti algatajatel oli kaks põhimõtet: võimalikult suure hulga inimeste projekti kaasamine ning kogu sekveneerimisinformatsiooni tasuta ja avalikuks kätte saadavaks tegemine. See parandab akadeemikute võimalusi kogutud andmeid lisauuringutes kasutada. Viimast põhimõtet on rakendanud ka järgnevad genoomi kaardistamise projektid [23].

3.1. 1000G

1000 genoomi (1000G) projekt loodi aastal 2008 ja selle eesmärk oli leida uuritavast valimist geneetilisi variante, mida esineb vähemalt 1% rahvastikus. 1000G loojate missioon oli koostada detailne inimese geneetilise variatsiooni kataloog, mida saaks kasutada laiem teadlaskond geneetiliste variantide assotsieerimisel uuritavate tunnustega. Uuringu jaoks kasutati erinevaid meetodeid, mille hulka kuulus ka täisgenoomide sekveneerimine. Projekti suure mahu tõttu jaotati projekt kolmeks etapiks.

Esimeses *piloot*-etapis sekveneeriti kahe perekonna triod, teiseks analüüsiti 179 indiviidi täisgenoomid ja kolmandaks vaadati 697 inimese tuhatkond juhuslikult valitud kodeerivat geeni. Pilootprojektiks kasutatud proovid saadi HapMap kollektsioonist [24]. Esimeses etapis uuriti kokku 1092 individuaali genoome. Valim koosneb 14 erinevast populatsioonist peamiselt Euroopa, Ida-Aasia, Aafrika ja Ameerika maailmajagudest. Kokku avastati 38 miljonit SNPsi, millele kõigile määrati ka genotüüp.

Teise etapis ei kogunud andmeid vaid tegeleti projekti tehnilise poolega. Kolmas etapp teostas ülejäänute indiviidide uuringu ja väljastas projekti finišeeritud andmeid. Valim koosnes 26 erinevatest populatsioonidest Aafrikast, Ameerikast , Ida-Aasiast, Euroopast ja Lõuna-Aasiast. Kokku analüüsiti mitme etapi vältel 2504 inimese geneetilist varieeruvust. Kokku avastati 84.7 miljon SNPsi, kus enamik (99%) on variandid, mille sagedus on vähemalt 1%. Järgnev joonis 2 näitab erinevate populatsioonide ‘unikaalsust’ võrreldes teiste rahvustega.

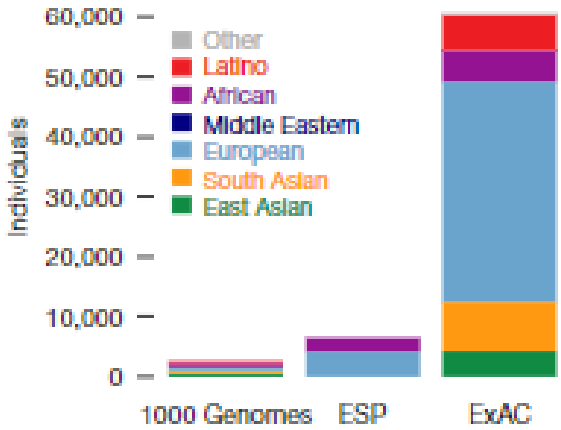


Joonis 2. 1000g populatsioonid maailmakaardil koos geneetilise sarnasuste sektordiagramm.

Iga sektor näitab ala kui palju ilmneb geneetilisi variante:

a) Tumehall ala –kõikides maailma osades
b) Helehall ala – maailma jagudes
c) Heledam värviline ala – enda maailmajaos
d) Tumedam värviline ala – enda populatsioonis

3.2. ExAC



Joonis 3. ExAC projekti valimi suurus võrreldes teiste sekveneerimise projektiga.

Värvidega on eristatud konkreetsete populatsioonide valmite arvu.

ExAC ehk ‘Eksoomiliitkonsortsiumiks’ (ingl *Exome Aggregation Consortium*) konsortsiumi eesmärk on ühendada mitmete suuremahuliste sekveneerimisprojektide andmed ühtseks kokkuvõtvaks andmebaasiks, mida saaks kasutada laiem teadlaskond. Konsortsiumi koondati 60 706 erineva etnilise taustaga indiviidi eksoomi. Joonis 3 näitlikustab Exaci kogutud andmemahtude suurust võrreldes teiste analoogiliste projektidega. Kokku leiti 7 404 909

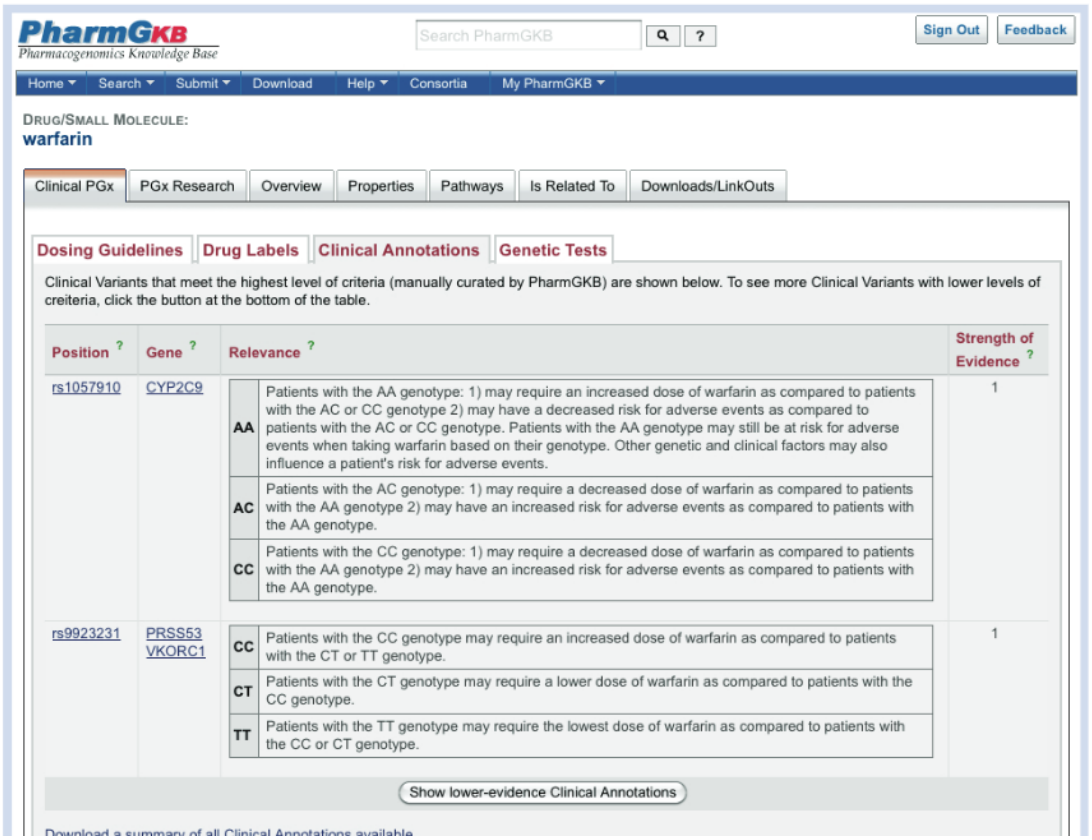
unikaalset varianti. Enamik variante on haruldased ja sagedusega alla 1%, umbes pooled variandid ilmnevad vaid korra ja 72% variantidest ei ilmne teistes tuntumates andmebaasides. Andmebaasi eesmärgiks on olla teadlastele vahendiks inimeste geneetilise variatsiooni mustrite tõlgendamisel ja kirjeldamisel. ExAC valimi suurus võimaldab hinnata geenide funktsionaalset profiili paremini kui seda seniste tööde põhjal on olnud võimalik teha [25].

4. Andmebaasid

Vabalt kättesaadavad andmebaasid katalogiseeritud geneetilise informatsiooniga on hindamatuks abiks teostatud uuringute tulemuste agregeerimisel ja korrastamisel. Üheks allikate rakenduseks on ka erinevate geneetiliste variantide ravimiseoste uurimine. Andmebaase saab kasutada meie töös leitavate alleelisageduste kontrolliks.

4.1. PharmGKB

PharmGKB on avalik ülemaailmne farmakogeneetiliste variantide andmebaas. Selle peamine eesmärk on leida ja levitada farmakogeneetikaga seotud informatsiooni. Eesmärgi teostamiseks on olemas internetis olev andmebaas, mida saab ka alla laadida. Veebilehel on nõuandeid teatud ravimite doseerimisega ja sisaldab geenide, ravimite ja haiguste kliinilisi märkusi. PharmGKB meeskond tegeleb ka valdkondlike andmebaaside analüüsimisega ja loob ülevaadet geneetiliste variantidega seotud teadusartiklitest.



Joonis 4. Näide PharmGKBs kirjeldatud ravim/alleel vaheliste seoste väljundist [22].

PharmGKB kasutajad saavad mugavalt informatsiooni variantide kohta teostades otsinguid ID-koodiga, geeni nimega või haiguse nimega. Kasutajale kuvatakse informatsiooni haiguste ja ravimite vahelisest seostest. Veebileht viitab teistele andmebaasidele ja teadusartiklitele, kus on lisainformatsiooni. Samuti sisaldab veebileht CPICi juhtnööre ehk aitab arstidel ravimite väljakirjutamisel geneetikat arvesse võtta [26]. Joonis 4. on näide PharmGKB väljundist haiguse ja ravimi vahelise seose kirjeldusest [22].

4.2. Ensembl

Ensembl on geneetilisi andmeid hõlmav portaal. See on Suurbritannias algatatud ühisprojekt Euroopa Bioinformaatika Instituudi, Euroopa Molekulaarbioloogia Laboritooriumi ja *Welcome Trust Sanger* Instituudi vahel. Projekt algas 1999. aastal eesmärgiga automaatselt geneetilist infot annoteerida, integreerida infot teiste bioloogiliste andmetega ja andmed vabalt avalikustada [27].

Ensembl projekt samuti tegeleb ka internetis oleva veebilehe haldamisega ja integreerib teisi genoomide ressursse. Veebilehel on ka kõikide oma liikide geneetiliste variantide tunnuseid ja võrdleva geneetika ressursse. Ensembl seob neid tunnuseid teiste internetis olevate vahenditega nagu PharmGKB, dbSNP (SNP üldine andmebaas), NCBI jne. Lisaks saavad kasutajad ka alla laadida Ensembli andmeid. Kasutajad leiavad otsingumootorit kasutades geneetilise variandi kohta erinevaid tunnuseid ja viiteid. Ensembl on üldotstarbeline mitmete organismide geneetilist infot agregeeriv portaal, kuid PharmGKB keskendub farmakogeneetiliste variantidega seonduvatele andmetele [28].

4.3. NCBI

Rahvusliku Biotehnoloogia Informatsiooni Keskus *(*ingl *National Center for Biotechnology Information aka NCBI)* loodi aastal 1988 Rahvusliku Terviseinstituudi poolt. Selle eesmärk on arendada uusi infotehnoloogilisi vahendid, et aidata paremini mõista põhilisi molekulaar- ja geneetilisi protsesse, mis mõjutavad tervist ja haigusi. NCBI ülesanne on luua automatiseeritud süsteeme, mis arhiveeriksid ja analüüsiksid informatsiooni molekulaarbioloogiast, bioloogilisest keemiast ja geneetikast.

Veebileht sisaldab väga palju erinevaid ressursse nagu geenide, nukleotiidide, kemikaalide jms andmebaasid. NCBI pakub ka erinevaid rakendusi bioloogiliste andmete analüüsimiseks ja erinevaid juhendeid andmete kasutamiseks. Veebilehel on ka teisi mitmeid kasulike ressursse nagu näiteks PubMed - 27 miljoni tsitaadi ja artiklite kataloog teemadel bioloogiline meditsiin, geneetika ja ravimid. Need teadusartiklid käsitlevad erinevaid seoseid geneetiliste variantide ja teatud ravimite vahel [29].

5. Statistilised meetodid populatsioonide eristamiseks

5.1. Peakomponentide analüüs

Peakomponentide analüüs (ingl *PCA aka Principal Components Analysis)* on statistiline meetod, mis vähendab kõrge dimensionaalsete andmestike dimensiooni koondades tunnustes leiduvat informatsiooni vähemate. PCA kasutab korreleeritud muutujaid, mida transformeeritakse väiksemaks arvuks mittekorreleeritud tunnusteks (peakomponentideks*).*

Esimene peakomponent kirjeldab esialgses andmestikus enim andmestikus leiduvast hajuvusest, kõik järgnevad peakomponendid on ortogonaalsed eelnevatega ning kirjeldavad maksimaalses ulatuses seni kirjeldamata andmetes leiduvast hajuvusest.

Näitlikustame peakomponentide analüüsi algoritmi. Olgu meil *m* dimensiooniga andmestik *A*, kus *m* on suvaline naturaalarv.

$$A = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

Maatriksi A iga rida on tunnuste kogum, ehk andmestiku esimest elementi iseloomustavad numbrid $x_{11}, x_{12}, \dots$ ja x_{1m}, teist elementi iseloomustab numbrid $x_{21}, x_{22}, \dots$ ja x_{2m} jne kuni viimase elemendi iseloomustavate numbriteni $x_{n1}, x_{n2}, \dots$ ja x_{nm}

PCA jaoks peab veergudest lahutama veergude aritmeetilise keskmise. Uue andmestiku tähis on A' ja luuakse järgnevalt:

$$A' = \begin{bmatrix} x_{11} - \bar{x_1} & x_{12} - \bar{x_2} & \dots & x_{1m} - \bar{x_m} \\ x_{21} - \bar{x_1} & x_{22} - \bar{x_2} & \dots & x_{2m} - \bar{x_m} \\ \dots & \dots & \dots & \dots \\ x_{n1} - \bar{x_1} & x_{n2} - \bar{x_2} & \dots & x_{nm} - \bar{x_m} \end{bmatrix}$$

Uue maatriksi esimest veergu tähistame tähega X_1, teist veergu tähistame tähega X_2 jne. kuni viimase veeruni, mis on tähistatud X_m tähega

$$X_1 = [x_{11} - \bar{x_1}, x_{21} - \bar{x_1}, \dots, x_{n1} - \bar{x_1}]$$

$$X_2 = [x_{12} - \bar{x_2}, x_{22} - \bar{x_2}, \dots, x_{n2} - \bar{x_2}]$$

$$\dots$$

$$X_m = [x_{1m} - \bar{x_m}, x_{2m} - \bar{x_m}, \dots, x_{nm} - \bar{x_m}]$$

Järgnevalt leitakse maatriksi, mis sisaldab kõikide veergude kovariatsioonid. Tähistame maatriksit tähega *covM* ja kovariatsiooni funktsiooniga (1).

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} \quad (1)$$

Maatriksit *covM* saadakse järgneva valemiga:

$$covM = \begin{pmatrix} cov(\mathrm{X}_1,\mathrm{X}_1) & cov(\mathrm{X}_1,\mathrm{X}_2) & \dots & cov(\mathrm{X}_1,\mathrm{X}_m) \\ cov(\mathrm{X}_2,\mathrm{X}_1) & cov(\mathrm{X}_2,\mathrm{X}_2) & \dots & cov(\mathrm{X}_2,\mathrm{X}_m) \\ \dots & \dots & \dots & \dots \\ cov(\mathrm{X}_m,\mathrm{X}_1) & cov(\mathrm{X}_m,\mathrm{X}_2) & \dots & cov(\mathrm{X}_m,\mathrm{X}_m) \end{pmatrix}$$

Saadud maatriksist leitakse omaväärtused λ valemiga:

$$\det(A-\lambda I) = 0$$

kus I on diagonaalne ühikmaatriks. Vastavalt determinandi definitsioonile saadakse omaväärtuste vektor:

$$Omaväärtus = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_m \end{bmatrix}$$

Omaväärtustest valitakse kõige suurem ja sellega arvutatakse omavektor (Herstein 1964). Peakomponendi mõiste ütleb, et omavektor, mis on saadud kõige suurema omaväärtusega, on andmestiku esimene peakomponent. Omavektorit arvutatakse järgneva valemiga:

$$(A-\lambda I)\mu = 0,$$

kus μ on omavektor. Esimene peakomponent tähistab kasutab suurimat omaväärtust ning kirjeldab suurima osa tunnuste hajuvusest. Kõik järgnevad peakomponendid kirjeldavad üha väiksemat osa seni kirjeldamata hajuvusest. Peakomponendid (PC) saame järgnevast maatrikskorrutisest.

$$PC = (\mu)^T \times (A')^T,$$

kus ^{T} tähistab transponeeritud maatriksi ja A' on (1) valemiga saadud maatriks [30].

5.2. F-statistik

Populatsioonide geneetikas ennustatakse mingi populatsiooni heterotsügootsust[3] kasutades F-statistikut (ingl *F-statistics aka Fixation index*). Seda kasutatakse mitmekesisuse uurimiseks alampopulatsioonide vahel [31]. F-statistiku väljund kirjeldab kahe rahvastiku geenide korrelatsiooni [32].

F-statistika arvutus teostatakse leides alampopulatsiooni geneetiline varieeruvus [33]. See tähistatakse sümboliga H_e ja saadakse valemiga:

$$He = 1 - (p^2 + q^2) \quad (2)$$

Siin on *p* geneetilise variandi sagedus vastavad populatsioonis ja *q=p-1*. Järgmisena on vaja leida sageduste aritmeetiline keskmine tähisega $\hat{p}$ valemiga:

$$\hat{p} = \frac{\sum_{j=1}^{i} p_j N_j}{\sum_{j=1}^{i} N_j},$$

kus N on alampopulatsiooni valimi suurus. Nüüd rakendatakse valemist (2) saadud tulemit alampopulatsioonide geneetilise varieeruvuse kogusumma leidmiseks. Seda tähistatakse sümboliga Hs ja saadakse valemiga:

$$Hs = \frac{\sum_{j=1}^{i} He_j N_j}{\sum_{j=1}^{i} N_j}$$

Edasi leitakse keskmine geneetiline varieeruvus vaadeldavates populatsioonides, mis on tähistatud tähega Ht ja avaldatud valemiga:

$$Ht = 1 - (\hat{p}^2 + \hat{q}^2)$$

Viimasena on F-statistik ehk populatsioonide geenide korrelatsiooni arvutus. Tähis on F_{st} ja valem on järgnev:

$$Fst = \frac{Ht - Hs}{Ht}$$

5.3. T-test

T-test on statistilise hüpoteesi test, millega kontrollitakse test statistiku pärinemist t-jaotusest. Seda kasutatakse enamasti kahe jaotuse keskmiste võrdlemiseks. T-testi väljunditeks on t-väärtus ja p-väärtus. T-väärtus kirjeldab kahe valimi erinevust. P-väärtus on tõenäosus, et valimi tulemused on juhuslikud kui kehtib nullhüpotees. Populaarne p-väärtuse piirmäära

[3] Heterotsügootsus – kahe erineva alleeli esinemine organismi kromosoomides

statistiliseks olulisuse piirmäär on 0.05 Kui p-väärtus on piirmäärast väiksem, siis need kaks valimit on statistiliselt erinevate keskmistega. Kui p-väärtus on suurem kui 0.05, siis nullhüpoteesi ümber ei lükata.

T-testi teostatakse järgnevate valemitega. Võrdleme kahte valimit *x* ja *y*. Valimi suurused on vastavalt n_x ja n_y

$$x = [x_1, x_2, \dots, x_{n_x}]$$
$$y = [y_1, y_{2,}, \dots, y_{n_y}]$$

Esiteks on vaja mõlema valimi aritmeetilist keskmist M_x ja M_y, mida saadakse järgnevate valemitega:

$$M_x = \frac{\sum_{i=1}^{n_x} x_i}{n_x}$$
$$M_y = \frac{\sum_{i=1}^{n_y} x_i}{n_y}$$

Teiseks leitakse valimite standardhälved tähistega s_x^2 ja s_y^2 järgnevate valemitega:

$$s_x^2 = \frac{\sum_{i=1}^{n_x}(x_i - M_x)^2}{n_x - 1}$$

$$s_y^2 = \frac{\sum_{i=1}^{n_y}(y_i - M_y)^2}{n_y - 1}$$

Viimaks saadakse t-testi teststatistik t-väärtus järgneva valemiga:

$$t = \frac{M_x - M_y}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_x^2}{n_x}}}$$

P-väärtust saadakse t-skoori konverteerimisega t-jaotuse tõenäosusfunktsiooni abil [34].

6. Praktiline töö

Nimetatud skeeme kasutatakse selleks, et uurida farmakogeneetiliste variantide profiili eestlaste andmetes. Antud peatükis kirjeldatakse praktilise töö andmeid, meetodeid ja tulemusi. Praktilist osa teostati statistilise programmeerimiskeelega R. Käesoleva bakalaureuse töö arvutused ja joonised koostati kasutades R keele kasutajaliidest R Studio [35]. Uuringus kasutatud R kood on kättesaadav GitHubi repositooriumist[4].

6.1. Andmed

Antud töös kasutati valimina VCF formaadis farmakogeneetiliste variantide sagedusandmeid Eesti Geenivaramu sekveneerimisprojektist, hollandlaste GoNL projektist ja UK10K projektist. Esialgsetes andmestikes olid eestlaste sagedusandmetes 3064 unikaalset geneetilist varianti, hollandlaste andmetes 3007 varianti ja UK10K andmetest 2952 varianti. Lisaks oli kasutatud 1000G viie rahvastiku andmebaasi, mis sisaldas 2770 unikaalset varianti ning ExAC andmestiku, mis sisaldas 3061 unikaalsete variantide sagedusandmeid. ExACi andmetest olid 792 geneetilist varianti kodeerivad. Kõikides andmebaasides olid lisaks sagedusandmetele teada ka kromosoom, positsioon kromosoomis, referentsalleel, alternatiivne alleel, kvaliteediskoor ja kvaliteedikontrolli tulemus. ExAC andmebaasist oli teada veel Ensembl andmebaasi rs-kood, alleel, geeni nimetus ja variandi ravimiga seotud omadus. Kõik uuritavad variandid olid farmakogeneetilised ehk nende kohta on varasemalt teada mingi mõju ravimi metaboliseerimise protsessis. Kolm andmestikku ühendati omavahel andmebaasiks, kus on iga geneetilise variandi rs-kood koos 3 erineva rahvastiku alleelisagedusega. Antud lõputöö praktilises osas nimetatakse uut andmebaasi 3 rahva populatsioonide andmebaasiks. Andmestikule on lisatud ka 1000G ja ExAC andmebaasidest saadud variantide sagedusandmed. Analüüsis kasutame SNPde ID-koode ning sagedusi Eesti, Hollandi ja Suurbritannia, 1000G (Aafrika, Ameerika, Ida-Aasia, Euroopa, Lõuna-Aasia) ja ExAC (Aafrika, Ladina-Ameerika, Ida-Aasia, Soome, Euroopa, Lõuna-Aasia ja muude) valimites. Kasutatakse F-statistikut üle kõikide uuritavate populatsioonide kodeerivate variantide (1274 SNPi). Lisaks on eraldi annoteeritud kõik VIP geenides asuvad variandid.

[4] https://github.com/Kennu76/genetical_data

Populatsioonid on tähistatud andmebaasidele spetsiifiliste lühenditega, mis märgivad geneetiliste andmete päritolu. Eesti, Hollandi ja Suurbritannia andmed on lühenditega EST_AF, NL_AF ja UK_AF. 1000G Aafrika, Ameerika, Ida-Aasia, Euroopa ja Lõuna-Aasia populatsioonid on tähistatud lühenditega AFR_MAF, AMR_MAF, EAS_MAF, EUR_MAF ja SAS_MAF. Eksoomliitkonsortsiumi ehk ExACi Aafrika, Ameerika, Soome, Euroopa (millest on välja arvatud soomlased), muude rahvaste ja Lõuna-Aasia populatsioonid on tähistatud vastavalt lühenditega ExAC_AFR_MAF, ExAC_AMR_MAF, ExAC_EAS_MAF, ExAC_FIN_MAF, ExAC_NFE_MAF, ExAC_OTH_MAF ja ExAC_SAS_MAF.

6.2. Ekstreemsete sagedustega variandid

Leiti igale rahvastikule variantide hulk, kus on vastavate variantide alleelisagedused teiste populatsioonidega võrreldes suurim või väikseim. Nende variantide võrdlus üle kõigi variantide näitab, kas ekstreemsed variandid on ühtlaselt jaotunud või on mõned populatsioonis üle- või alaesindatud. Kõrvalekalded on indikatsiooniks erinevatest valikulistest survetest. Kõikide populatsioonide jaoks tuvastatakse absoluutse sageduse poolest maksimaalsed ja minimaalsed variandid.

Eestlaste ekstreemsete variantide seas leiti ka unikaalsemaid variante, mille sagedus on märgatavalt suurem või märgatavalt väiksem kui ülejäänud maailma populatsioonides. Olgu geneetilise variandi sagedused tähistatud sümboliga *var*.

$$var = \begin{bmatrix} NL_{AF}, EST_{AF}, UK_{AF}, AFR_{AF}, AMR_{AF}, EAS_{AF}, EUR_{AF}, SAS_{AF} \\ , ExAC\_AFR\_MAF_{AF}, ExAC\_AMR\_MAF_{AF}, ExAC\_EAS\_MAF_{AF} \\ , ExAC\_FIN\_MAF_{AF}, ExAC\_NFE\_MAF_{AF}, ExAC\_OTH\_MAF_{AF}, ExAC\_SAS\_MAF_{AF} \end{bmatrix}$$

Siin on NL_{AF} variandi sagedus hollandlastel, EST_{AF} on variandi sagedus eestlastel, UK_{AF} on sagedus Suurbritannia elanikkudel, AFR_{AF} on sagedus 1000G andmebaasi aafriklastel, AMR_{AF} on sagedus 1000G andmebaasi ameeriklastel, EAS_{AF} on sagedus 1000G andmebaasi ida-asiaatidel, EUR_{AF} on sagedus 1000G andmebaasi eurooplastel, SAS_{AF} on sagedus 1000G andmebaasi lõuna-asiaatidel, ExAC_AFR_MAF_{AF} on sagedus ExAC andmebaasi aafriklastel, ExAC_AMR_MAF_{AF} on sagedus ExAC andmebaasi ameeriklastel, ExAC_EAS_MAF_{AF} on sagedus ExAC andmebaasi ida-asiaatidel, ExAC_FIN_MAF_{AF} on sagedus ExAC andmebaasi

soomlastel, ExAC_NFE_MAF_{AF} on sagedus ExAC andmebaasi eurooplastel (ilma soomlasteta), ExAC_OTH_MAF_{AF} on sagedus ExAC andmebaasi muudel rahvastel, ExAC_SAS_MAF_{AF} on sagedus ExAC andmebaasi lõuna-asiaatidel.

Järgnevalt leitakse eestlaste andmetes ekstreemsetele variantidele suhteline kaugus ülejäänud populatsioonide sagedustest. Selleks eemaldatakse EST_{AF} valimist ning leitakse uuest valimist kõige suurem sagedus *max(var)*. Seejärel leitakse suhteline kaugus, omistades variandile protsentuaalne kaugus järgmisest:

$$\%Diff = \frac{EST_{AF} - \max(var)}{\max(var)}$$

Ekstremaalselt väikese eestlaste sageduse korral leitakse uuest valimist kõige väiksem sagedus *min(var)* ja omistatakse variandile protsentuaalne vahe analoogilise valemiga:

$$\%Diff = \frac{\min(var) - EST_{AF}}{\min(\mathrm{var})}$$

%Diff leitakse kõikidele eestlaste seas ekstreemsete variantidele. See mõõt aitab prioritiseerida variantide uurimisjärjekorda.

6.3. Peakomponentide analüüs

Farmakogeneetiliste variantide omavahelise sarnasuse ja populatsioonide omavahelise sarnasuse uurimiseks kasutame peakomponentide analüüsi. Uuritakse SNPide omavahelist lähedust üle erinevate populatsioonide. Kasutati R programmeerimiskeele funktsiooni nimega *prcomp*, mille sisend on reaalarvuliste väärtustega maatriks. Selle tulemusena leiame variantide varieeruvust kirjeldavad peakomponendid.

Esimeses PCA etapis uurime variantide lähedust inglaste, hollandlaste ja eestlaste alleelisageduste põhjal. Analüüsist saadud esimest ja teist peakomponenti kasutatakse selleks, et kirjeldada variantide omavahelisi erinevusi. Samasugune uuring tehakse ka eestlaste ja 1000G rahvaste alleelisagedustele. Järgmisena kombineeriti erinevaid andmebaase omavahel. Teostati uuring andmetega, milles on ühendatud 3 rahva populatsioonid, 1000G rahvad ja

ExAC rahvad, et avastada nende populatsioonide vahelisi erinevusi või sarnasusi. Variandid grupeeritakse maksimaalse alleelisagedusega populatsioonide järgi ning valitud gruppides võrreldakse peakomponentide keskmisi väärtusi, kasutades t-testi.

Kolmandas etapis uuriti populatsioonide omavahelisi erinevusi kasutades farmakogeneetiliste variantide sagedusandmeid. Sellega transformeeritakse antud maatriksit, nii et andmemaatriksi ridadeks on populatsiooniandmed üle variantide, mitte geneetilised variandid üle populatsioonide.

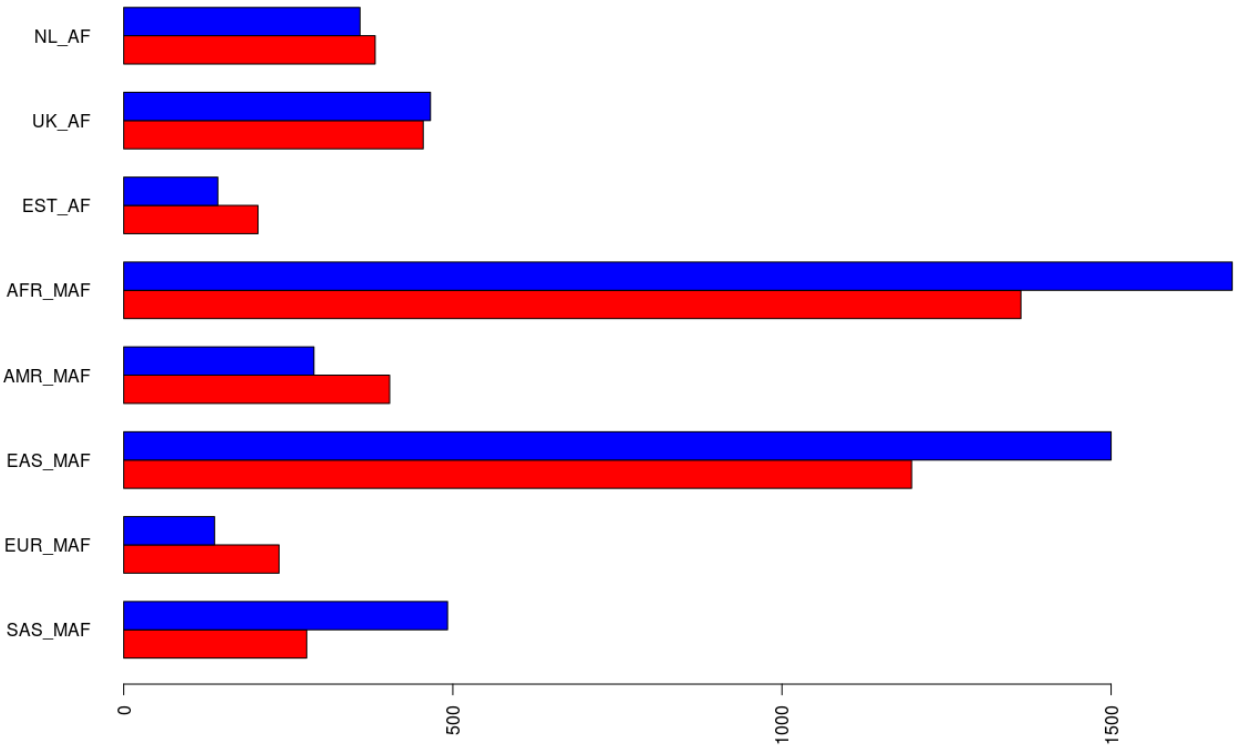
Viimaks uuriti erinevaid gruppe rakendades t-testi. Võrreldi omavahel kodeerivate variantide gruppi ja mitte-kodeerivate variantide gruppi esimesi peakomponente. Uuriti neid eristavaid gruppe ja teostati t-testi kahest analüüsist saadud tulemite vahel. Samamoodi võrreldi ka VIP variantide gruppi esimest peakomponenti mitte-VIP variantidega.

6.4. F-statistiku rakendamine

Populatsioonide paariviisilise heterogeensuse hindamiseks farmakogeneetiliste variantide põhjal kasutati F-statistikut. Suurbritannia, Hollandi ja Eesti elanike alleelisagedused ning valimisuurused saadi täisgenoomide andmetest [36]. 1000G ja ExAC andmete jaotust populatsioonide kaupa saadi algallikatest [37]. Saadud tulemustega loodi R Studio funktsiooni „soojuskaardi“ (ingl *heatmap*). Tulemused kirjeldavad populatsioonide omavahelist sarnasust.

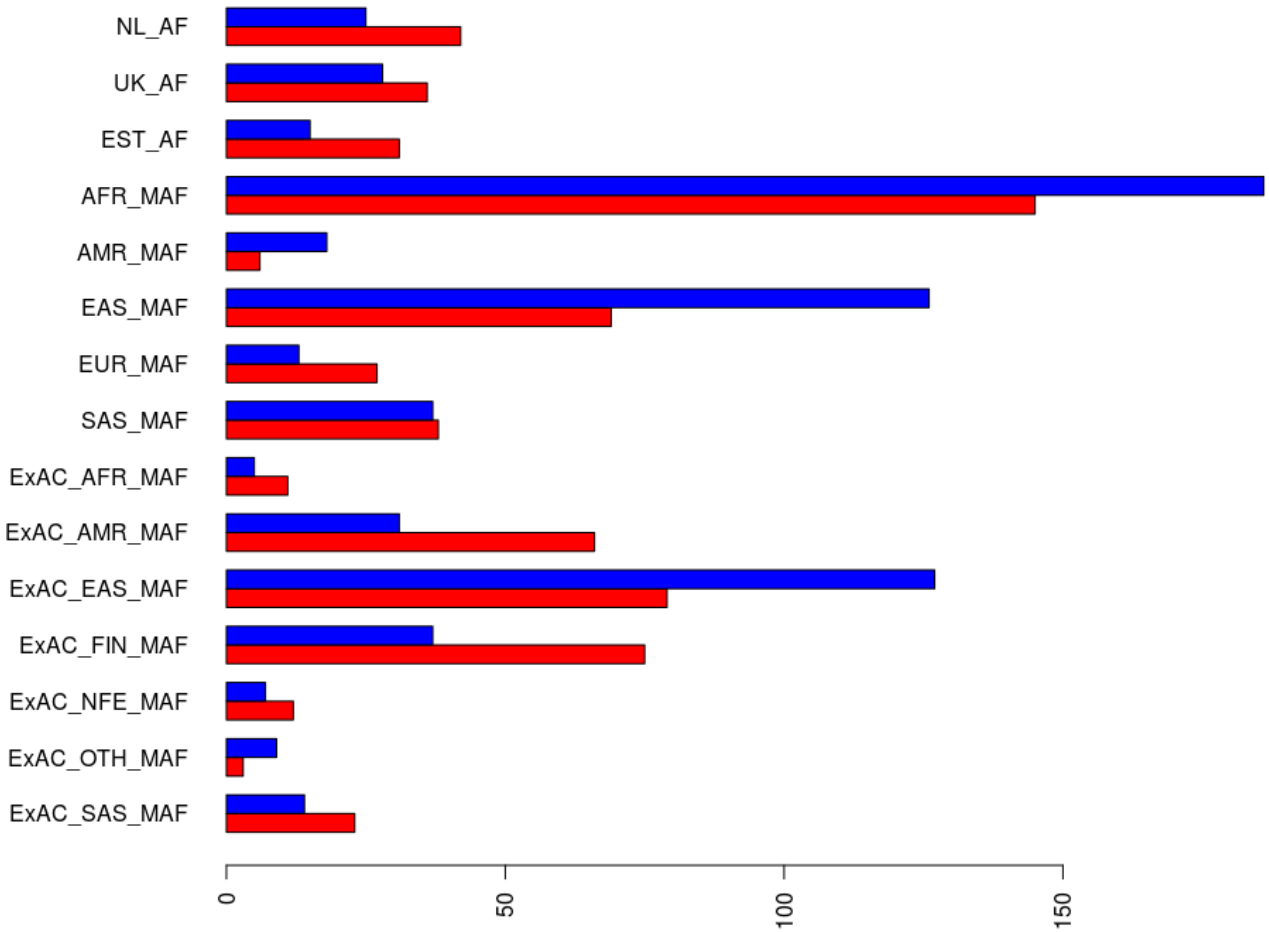
7. Praktilise töö tulemused

7.1. Ekstreemsed variandid



Joonis 5. Mitte-kodeerivate farmakogeneetiliste variantide ekstremaalsete sageduste jaotumine 1000G ja kolmes Euroopa populatsioonis (NLGo, UK10K, Eesti). Punasega on tähistatud variandid, mille sagedus on teistega võrreldes suurim, sinisega on tähistatud minimaalse sagedusega variandid.

Joonisel 5. on kolme rahva populatsioonid ja 1000G populatsioonide farmakogeneetilisete variantide minimaalsed ja maksimaalsed sagedushulgad. Märgatakse, et aafriklastel ja ida-asiaatidel on suur hulk ekstreemseid variante.



Joonis 6. Kodeerivate farmakogeneetiliste variantide ekstremaalsete sageduste jaotumine Exac, 1000G ja kolme Euroopa populatsiooni (NLGo, UK10K, Eesti) vahel. Punasega on tähistatud variandid, mille sagedus on teistega võrreldes suurim, sinisega on tähistatud minimaalse sagedusega variandid.

Joonis 6. kirjeldab ekstreemseid variante üle kõigi kodeerivate variantide. Siin on aafriklased (AFR_MAF) ja ida-asiaadid (EAS_MAF ja ExAC_EAS_MAF) teistest erinevamad, kuid ka soomlastel (ExAC_FIN_MAF) on palju ekstreemseid variante.

Tabel 2. Ekstremaalsete sagedustega farmakogeneetilised variandid Eesti andmetes.

Kõik ekstremaalsed variandid					VIP variandid				
refsnp_id	EST_AF	UK_AF	NL_AF	VIP	refsnp_id	EST_AF	UK_AF	NL_AF	VIP
rs1138272	0.0983058	0.079476	0.083	1	rs1138272	0.0983058	0.079476	0.083	1
rs12208357	0.0955882	0.073129	0.071	1	rs12208357	0.0955882	0.073129	0.071	1
rs12819210	0.266266	0.184607	0.151	0	rs1799929	0.444296	0.429516	0.441	1
rs17602729	0.171346	0.130918	0.15	0	rs34059508	0.0282977	0.023407	0.014	1
rs2120825	0.165775	0.099841	0.1	0	rs4149056	0.215241	0.14996	0.162	1
rs3219484	0.0989305	0.076303	0.061	0	rs4846051	1	0.999207	0.998	1
rs34059508	0.0282977	0.023407	0.014	1	rs6018	0.113636	0.05369	0.043	1
rs6018	0.113636	0.05369	0.043	1	rs8192709	0.0811052	0.044433	0.068	1
rs72551330	0.0245098	0.014282	0.007014	0	rs1208	0.549911	0.576303	0.558	1
rs8192709	0.0811052	0.044433	0.068	1	rs2066853	0.0964795	0.110553	0.11	1
rs10380	0.0577094	0.093097	0.093	0	rs2306283	0.39082	0.400026	0.407	1
rs162036	0.0637255	0.116504	0.108	0	rs4149032	0.240196	0.346205	0.334	1
rs2071554	0.0298574	0.036631	0.049	0	rs701265	0.113636	0.152605	0.14	1
rs2242047	0.00735294	0.013885	0.012	0					
rs4149032	0.240196	0.346205	0.334	1					
rs4149032	0.240196	0.346205	0.334	0					
rs6907567	0.157086	0.228775	0.208	0					
rs701265	0.113636	0.152605	0.14	0					
rs701265	0.113636	0.152605	0.14	1					
rs714368	0.156863	0.228379	0.209	0					

Eestlaste andmetes olevad unikaalsed variandid on välja toodud tabelis 2, kus on vastavalt ka variandi alleelisagedustega Eestis, Hollandis ja Suurbritannias ning variandi rs-kood. Vasakul pool on eesti valimi esimesed 10 suurima protsentuaalse kaugustega ekstreemset varianti ja paremal pool on ekstreemsed variandid VIP geenides. Nende variantide uurimiseks kasutati interneti olevaid andmebaasi PharmGKB. Kõik unikaalsed VIP variandid mõjutasid vähemalt ühe ravimi metabolismi. Kõige suuremat huvi pakkusid viis geneetilist varianti, mis kuuluvad PharmGKB andmebaasis VIP variandi all, mis on vähemalt 3 taseme toksilisuse (ingl *Toxicity*) annotatsiooniga.

Esimene huvitav variant oli koodiga rs1799929 ja asub geeni *NAT2* . Selle unikaalse variandi sagedus eestlastel on 44.4%, suurbritannias on sagedus 43% ja hollandlastel on sagedus 44.1%. 1000G ja ExAC populatsioonide sagedus on keskmiselt 29.6%. Antud variandi kandjatel on suurem risk tuberkuloosi vastaste ravimite tarbimise korral tüsistada maks [38].

Teine huvitav variant oli koodiga rs34059508 ja asub geenis *SLC22A1* . Selle unikaalse variandi sagedus eestlastel on 2.8%, Suurbritannias on sagedus 2.3% ja hollandlastel on sagedus 1.4%. 1000G ja ExAC populatsioonide sagedus on keskmiselt 0.7%. Antud variandi kandjatest laste organismis püsib morfiin kauem kui mittekandjatel [39].

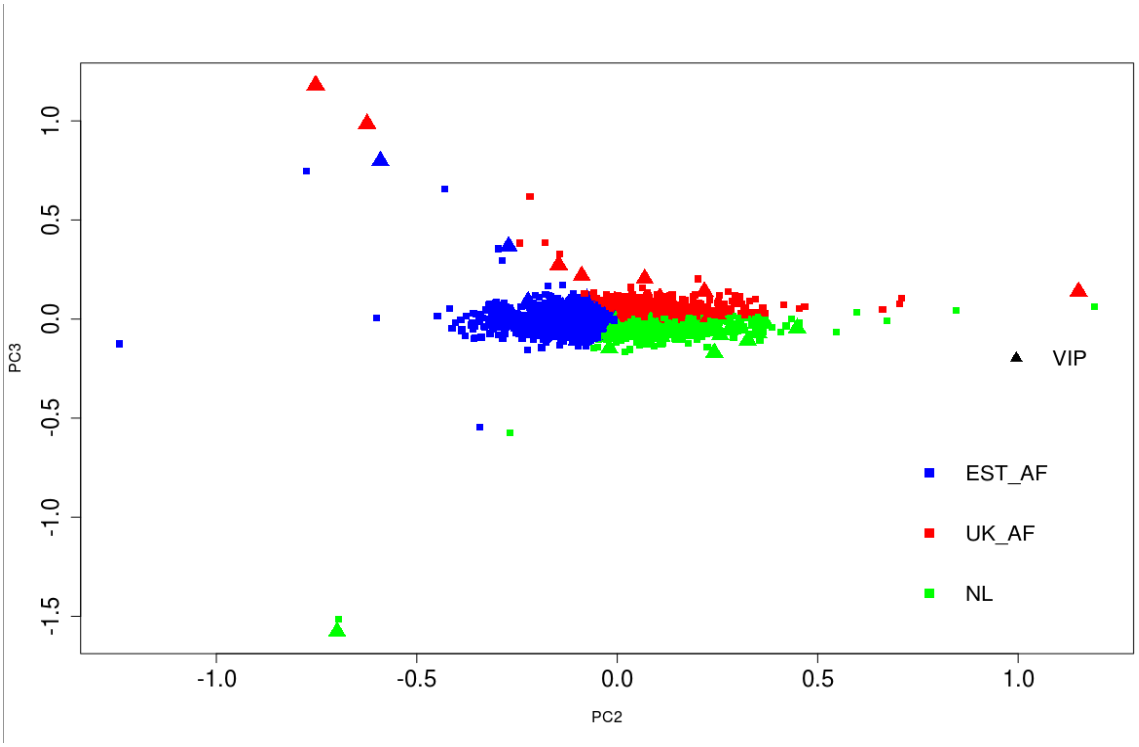
Kolmas märkimisväärne variant, koodiga rs2306283, asub geenis *SLCO1B1*. Selle unikaalse variandi sagedus eestlastel on 39%, Suurbritannias on sagedus 40.7% ja hollandlastel on sagedus 40%. 1000G ja ExAC populatsioonide sagedus on keskmiselt 56.3%. Antud variandi kandjatel on statiinide põhiliste ravimite tarbimisel tõenäolisemalt kõrgemad kreatiinkinaasi tase ja suurem risk ravi talumatuseks [40].

Järgmine huvipakkuv variant oli koodiga rs4846051 asub geenis *MTHFR*. Selle unikaalse variandi sagedus eestlastel on 100%, Suurbritannias on sagedus 99.9% ja hollandlastel on sagedus 99.8%. 1000G ja ExAC populatsioonide sagedus on keskmiselt 92.8%. Antud variandi kandjatel on kõrgem ravimist tulenev toksilisuse tase kui nad tarbivad metotreksaati reumatoidartriidi tõttu [41].

Geneetiline variant rs4149056 esineb geenis *SLCO1B1*, selle geeni *521TC* genotüübis. Selle alleelisagedus eestlastel on 21.5%, suurbritannia inimestel 15%, hollandlastel 16.2% ja teistes populatsioonidel on see keskmiselt 11%. Variandi kandjad omastavad erinevalt statiinide põhiseid ravimeid. Richard Ho uuring tõestas, et selle variandi kandjatel on vaja manustada väiksem annus [42]. Teine uuring avastas, et variandi kandjatel on kõrgem risk statiinraviga seotud müopaatia[5] tekkeks. Sama uuring leidis ka seda, et simvastatiin alandab enam kolesterooli [43].

[5] Lihaspõletik

7.2. Peakomponendi analüüside tulemused



Joonis 6. Eesti, NLGo ja UK10K populatsioonide peakomponendi analüüs. Iga punkt kirjeldab ühte geneetilist variant, suur kolmnurk on farmakogeneetiline variant VIP geenis. Sinised punktid on eestlaste-, punased on UK10K andmestiku ja rohelised on NLGo populatsioonis maksimaalse sagedusega variandid

3 rahva populatsioonide peakomponentide analüüsi tulemus on kajastatud joonisel 6. Joonisel olevad punktid on geneetilised variandid, mis on värvitud vastavalt populatsioonile, milles on alleelisagedus teistest populatsioonidega võrreldes suurim. Sinised punktid on eestlaste (EST_AF) variandid, rohelised on NLGo (NL_AF) variandid ja punased on UK10K (UK_AF) variandid. Suured kolmnurgad tähistavad geneetilisi variante VIP geenides. Joonisel on teise ja kolmanda peakomponendi skoorid. Iga peakomponendi skoor on lineaarne kombinatsioon individuaalsetest tunnustest ja nende kaaludest. Joonisel 6 kirjeldatud teise peakomponendi skoor avaldub:

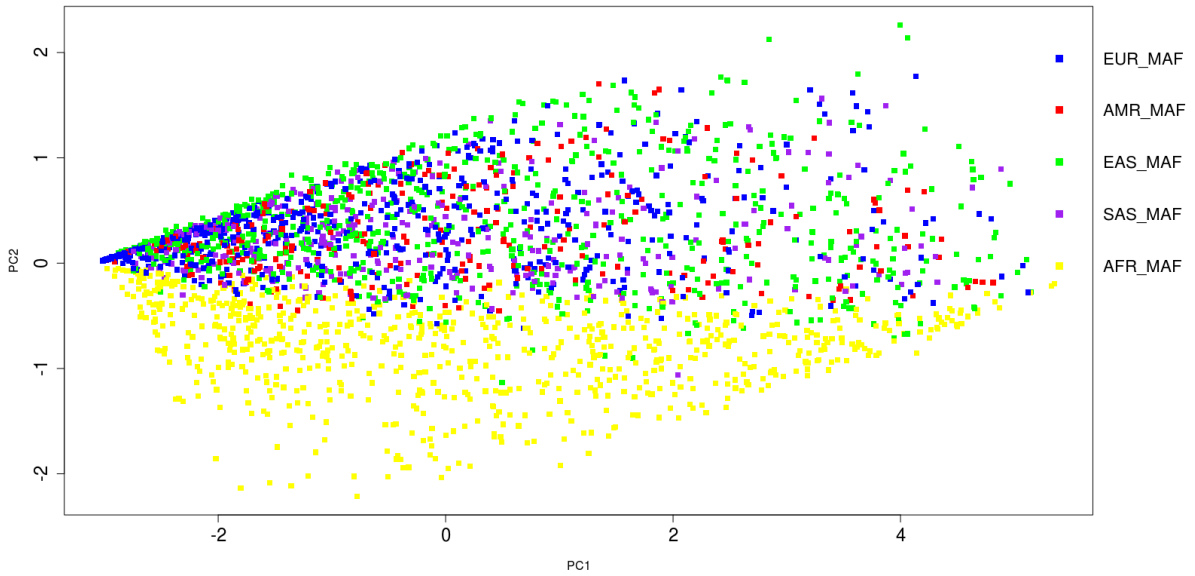
$$PC2 = 0.45 * NL_{AF} + 0.36 * UK_{AF} - 0.82 * EST\_AF$$

ja kolmanda peakomponendi skoor on:

$$PC3 = -0.68 * NL_{AF} + 0.73 * UK_{AF} - 0.05 * EST\_AF$$

Joonisest on näha, et 3 populatsiooni eristuvad üksteisest. Nähakse, et teatud VIP variandid on kõikidel populatsioonidel eristavamad kui tavalised variandid. Näiteks eristuvad 3 VIP

varianti, mis asuvad joonisel kõige üleval. Need on variandid koodidega rs7438284, rs7439366 ja rs1801253. Esimene ja teine variant on suhteliselt sarnased, kuna nad mõlemad asuvad *UGT2B7* geenis ja mõlemal on sagedus Eestis 44.3%, Suurbritannias 45.2% ja Hollandis 0.9% ja 8%. Uuringud näitavad, et nende variandi kandjad võtavad paremini vastu ravimit nimega lorazepam, mida kasutatakse ärevushäiret vastu ja valproatet, millega ravitakse epilepsiat [44]. Kolmas variant asub geenis *ADRB1* ja selle sagedus on Eestis 75.9%, Suurbritannias 73.8% ja Hollandis 43.8%. Antud geneetilise variandi kandjad omastavad ravimit metoprololi paremini ehk tema mõju vererõhu madaldamiseks on suurem [45].



Joonis 7. 1000G populatsioonide peakomponendi analüüs.

Kolmnurgad tähistavad VIP variante. Populatsioonide ekstreemsed variandid on märgitud järgnevalt: sinised punktid on eurooplaste ekstreemsed variandid, punased on ameeriklaste ekstreemsed variandid, ida-asiaatidel on roheline värv, lõuna-asiaatidel on violetne ja aafriklastel kollane värvus

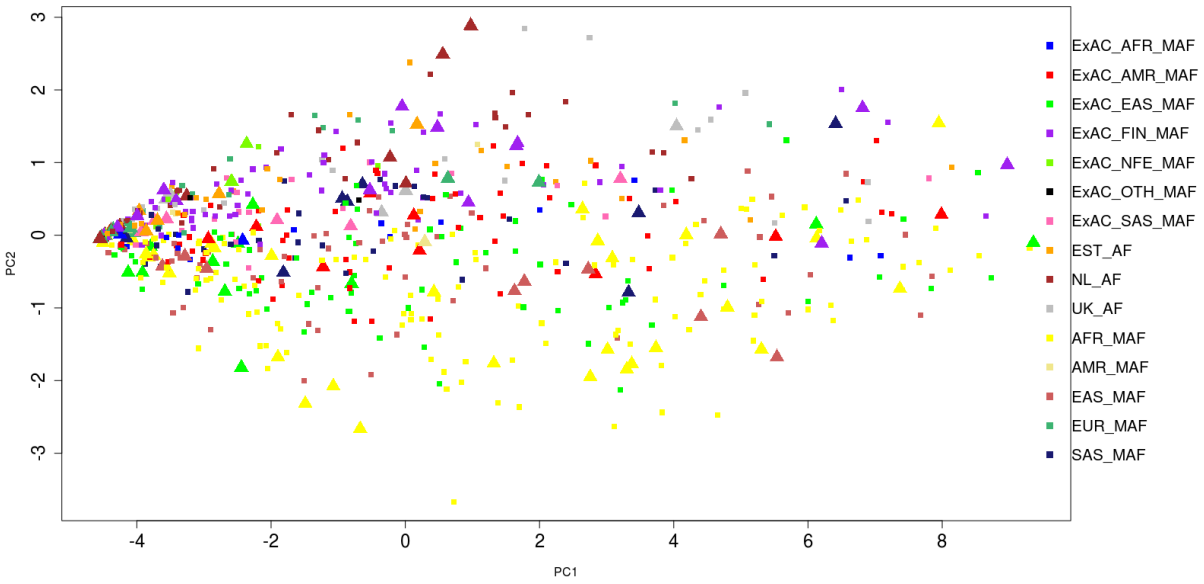
Joonis 7. kujutab endas 1000G rahvaste alleelisageduste andmetega, kus esineb märgatav ülekate VIP ja mitte-VIP variantide vahel. Järjekordselt on joonisel olevad punktid värvitud vastavalt sellele, millises populatsioonis on selle alleelisagedus kõige suurem. Sinised punktid on eurooplaste (EUR_MAF) variandid, punased on ameeriklaste (AMR_MAF) variandid, rohelised on ida-asiaatide (EAS_MAF) variandi, violetsed on lõuna-asiaatide (SAS_MAF) variandid ja kollased of aafriklaste (AFR_MAF) variandid.. Esimese peakomponendi skoor on:

$$PC1 = 0.38 * AFR\_MAF + 0.48 * AMR\_MAF + 0.44 * EAS\_MAF + 0.46 * EUR\_MAF + 0.47 * SAS\_MAF$$

ja teise peakomponendi skoor on:

$$PC2 = 0.91 * AFR\_MAF + 0.41 * AMR\_MAF + 0.24 * EAS\_MAF + 0.22 * EUR\_MAF + 0.16 * SAS\_MAF$$

Uuringus kasutatud populatsioonid võrreldi ameeriklastega ja eurooplastega kuna viimased on joonistest kõige ühtlasemalt jaotunud ehk need populatsioonid on kõige sarnasemad ülejäänud populatsioonidega. Teostati t-teste aafriklaste, eurooplaste ja ameeriklaste ekstreemsete variantide teise peakomponenti skooride vahel. Aafriklased on teistest populatsioonidest rohkem eraldatud ehk nemad on geneetiliselt omapärasemad kui teised võrreldud populatsioonid. Aafriklaste ja eurooplaste t-väärtus oli 48.556 ning aafriklaste ja ameeriklaste t-väärtus oli 35.564. Mõlema p-väärtus oli väiksem kui $1*10^{-16}$, mis tähendab, et nii aafriklaste ja eurooplaste kui aafriklaste ja ameeriklaste peakomponentide keskmiste vahel esineb erinevus. Ameeriklaste ja eurooplaste t-väärtus oli -1.0747 ja p-väärtus oli 0.2828 ehk nende kahe populatsiooni vahel ei ilme geneetilisi erinevusi. Sama kehtib ka teiste populatsioonide kombinatsioonidega, kus ei ole aafriklasi näiteks ameeriklaste ja lõuna-asiaatide t-väärtus oli 0.83411 ja p-väärtus oli 0.4045



Joonis 8. 3 populatsioonide, 1000G populatsioonide ja ExAC populatsioonide ühisele andmebaasile teostatud peakomponendi analüüs. Igal populatsiooni maksimaalse sagedusega variandid on tähistatud erineva värvusega, mis on nähtav joonise legendist.

Joonis 8. on peakomponentide analüüsi tulemuste kujutis, mis on koostatud kõikide populatsioonide andmetega. Sellel joonisel on märgitud nii 3 rahvastiku andmed, 1000G rahvaste andmed kui ka ExAC rahvaste andmed. Kõik joonisel olevad geneetilised variandid on kodeerivad. Esimese peakomponendi skoor on:

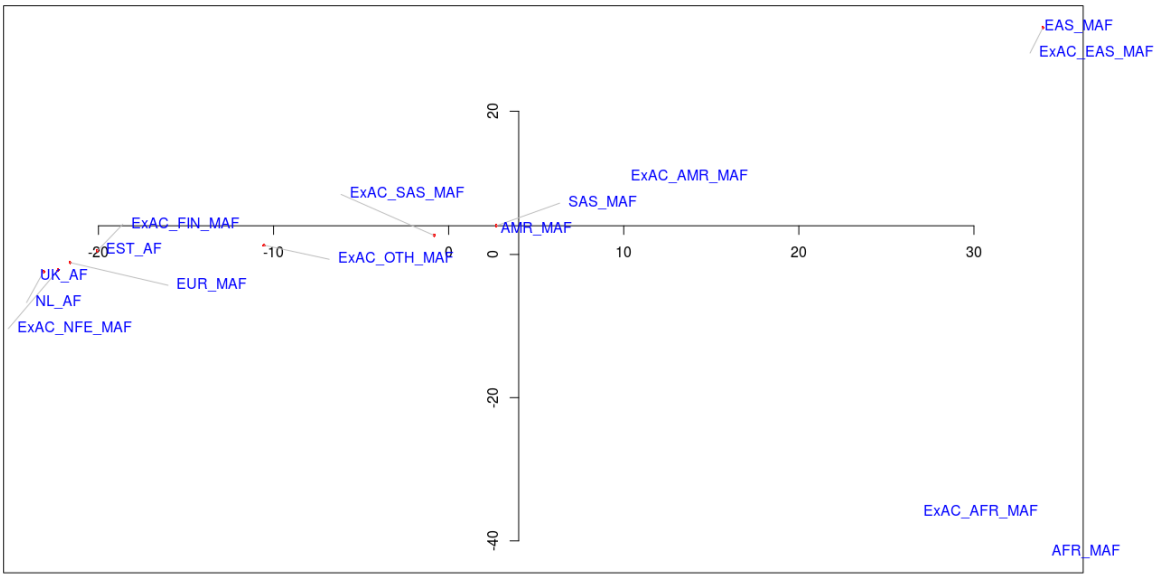
$$PC1 = 0.26 * EST\_AF + 0.26 * UK\_AF + 0.26 * NL\_AF + 0.22 * AFR\_MAF + 0.24 * EAS\_MAF + 0.26 * EUR\_MAF + 0.26 * SAS\_MAF + 0.24 * ExAC\_AFR\_MAF + 0.26 * ExAC\_AMR\_MAF + 0.24 * ExAC\_EAS\_MAF + 0.26 * ExAC\_FIN\_MAF + 0.27 * ExAC\_NFE\_MAF + 0.27 * ExAC\_OTH\_MAF + 0.27 * ExAC\_SAS\_MAF$$

ja teise peakomponendi skoor on:

$$PC2 = 0.22 * EST\_AF + 0.25 * UK\_AF + 0.26 * NL\_AF + -0.46 * AFR\_MAF + -0.03 * AMR\_MAF + -0.35 * EAS\_MAF + 0.24 * EUR\_MAF + -0.05 * SAS\_MAF + -0.39 * ExAC\_AFR\_MAF + -0.08 * ExAC\_AMR\_MAF + -0.34 * ExAC\_EAS\_MAF + 0.22 * ExAC\_FIN\_MAF + 0.24 * ExAC\_NFE\_MAF + 0.11 * ExAC\_OTH\_MAF + -0.006 * ExAC\_SAS\_MAF$$

,

Järeldatakse, et soomlased (ExAC_FIN_MAF) ja aafriklased (ExAC_AFR_MAF) on omapärased, kuna VIP variandid kogunevad kobarasse erinevates kohtades joonises. Seda kontrolliti ka erinevate t-testidega. Võrreldi eelnevalt mainitud populatsioone ameeriklastega esimest peakomponenti kasutades. Soomlaste ja ameeriklaste geneetilised andmed erinesid, kuna nende t-väärtus oli 2.228 ja p-väärtus oli 0.02714. Aafriklaste ja ameeriklaste geneetilised andmed erinesid, kuna nende t-väärtus oli -2.3033 ja p-väärtus oli 0.02183. Teiste populatsioonide vahelised t-testide tulemustes ei kajastatud erinevust. Ameeriklaste ja ida-asiaatide t-väärtus oli -0.73041 ja p-väärtus oli 0.4659. Samamoodi ei ilmnenud erinevus ka eurooplaste ja lõuna-asiaatide vahel, mille t-väärtus oli -1.5485 ja p-väärtus oli 0.1247. Järelikult on soomlaste, aafriklaste ja hollandlaste geneetilised variandid omapärasemad, kuna t-testid näitasid, et need populatsioonid erinevad ülejäänud populatsioonidest. Nimetatud teiste populatsioonide enda vahel ei esinenud erinevusi.



Joonis 9. Kõikide populatsioonide ühisele andmebaasile (Exac, 1000G, NLGo, UK10K, Eesti) teostatud peakomponendi analüüs.

Joonisel 9 on näha populatsioonide kahe esimese peakomponendi väärtused üle kõigi kodeerivate farmakogeneetiliste variantide. Populatsioonid paiknevad loogiliselt vastavalt geograafilisele kaugusele originaalsetel asualadel. Euroopa (EST_AF, UK_AF, NL_AF, ExAC_NFE_MAF, EUR_MAF, ExAC_FIN_MAF) rahvad on omavahel sarnased. Järjekordselt märgatakse, et Ida-Aasia (EAS_MAF, ExAC_EAS_MAF) rahvad on äärmiselt erinevad, aga Ameerika (AMR_MAF, ExAC_AMR_MAF) ja Lõuna-Aasia (SAS_MAF, ExAC_SAS_MAF) rahvad on üksteisele sarnased. Lisaks sellele on ka Aafrika (AFR_MAF, ExAC_AFR_MAF) rahvad ülejäänutest hästi erinevad. Erinevus ja sarnasus on joonisel kirjeldatud tasandil olevate populatsioonide omavahelisest lähedusest. Lähedased populatsioonid on sarnased ja üksteisest kaugel olevad populatsioonid on erinevad. Esimese peakomponendi skoor oli:

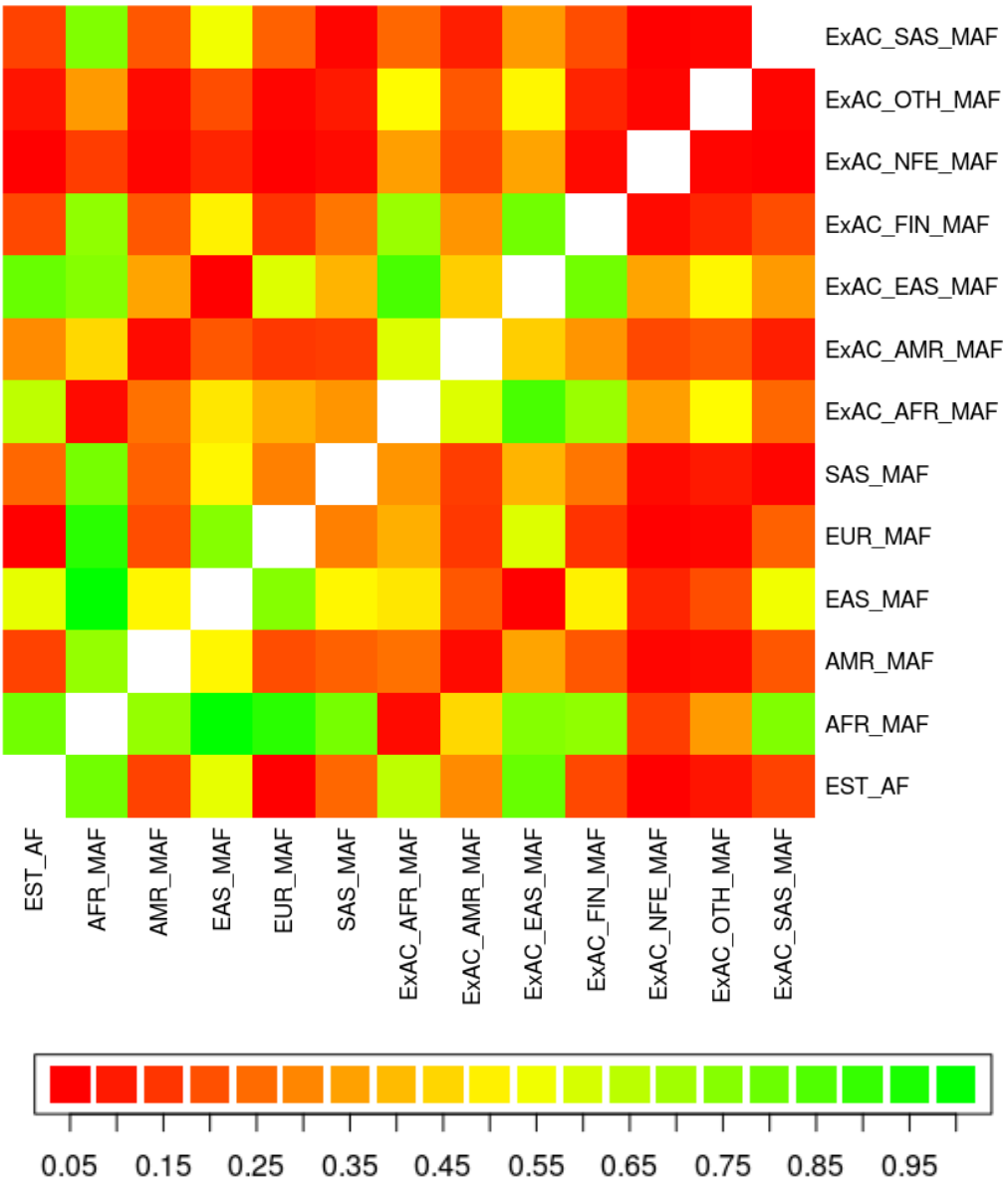
$$\mathrm{PC1} = -20.07 * \mathrm{EST\_AF} + -22.27 * \mathrm{UK\_AF} + -23.13 * \mathrm{NL\_AF} + 33.95 * \mathrm{AFR\_MAF} + 33.55 * \mathrm{EAS\_MAF} + -21.61 * \mathrm{EUR\_MAF} + 2.72 * \mathrm{SAS\_MAF} + 26.60 * \mathrm{ExAC\_AFR\_MAF} + 7.60 * \mathrm{ExAC\_AMR\_MAF} + 33.94 * \mathrm{ExAC\_EAS\_MAF} + -20.06 * \mathrm{ExAC\_FIN\_MAF} + -22.29 * \mathrm{ExAC\_NFE\_MAF} + -10.55 * \mathrm{ExAC\_OTH\_MAF} + -0.82 * \mathrm{ExAC\_SAS\_MAF}$$

Teise peakomponendi skoor oli:

$$\mathrm{PC2} = 0.59 * \mathrm{EST\_AF} + -2.02 * \mathrm{UK\_AF} + -2.39 * \mathrm{NL\_AF} + -41.52 * \mathrm{AFR\_MAF} + 3.52 * \mathrm{AMR_{MAF}} + 31.79 * \mathrm{EAS_{MAF}} + -1.14 * \mathrm{EUR_{MAF}} + 4.01 * \mathrm{SAS\_MAF} + -35.98 * \mathrm{ExAC\_AFR\_MAF} + 9.2 * \mathrm{ExAC\_AMR\_MAF} + 31.67 * \mathrm{ExAC\_EAS\_MAF} + 0.49 * \mathrm{ExAC\_FIN\_MAF} + -2.18 * \mathrm{ExAC\_NFE\_MAF} + 1.28 * \mathrm{ExAC\_OTH\_MAF} + 2.65 * \mathrm{ExAC\_SAS\_MAF}$$

T-testid, mis uurisid esimese peakomponendi vahelisi erinevusi VIP geenidesse kuulumise põhjal eristuvates kodeerivates variantides; kodeerivates ning mittekodeerivates ja ka VIP geenidesse kuuluvates kodeerivates ning mittekodeerivates variantides näitasid gruppide vaheliste erinevuste puudumist (p-väärtus 1).

7.3. F-statistika meetodi tulemused



Joonis 10. Soojuskaart populatsioonide (ExAC, 1000G, Eesti) heterogeensusest F-statistiku põhjal. Punane värv näitab, et kaks populatsiooni on omavahel hästi sarnased. Kollane värvus näitab, et populatsioonide omavaheline erinevus või sarnasus ei ole eripärane Roheline värvus näitab, et kaks rahvastik on omavahel geneetiliselt hästi erinevad. Väiksem F-statistiku väärtus kirjeldab vähemat erinevust kahe populatsiooni vahel.

Paariviisilised populatsioonide heterogeensused kirjeldatuna F-statistikuga on kujutatud joonisel 10. Sellel joonisel käsitletakse Eesti, ExACi ja 1000G populatsioonide variantide sagedusi. Siin märgatakse järjekordselt, et aafrika (AFR_MAF, ExAC_AFR_MAF) ja ida-aasia populatsioonid (EAS_MAF, ExAC_EAS_MAF) on kõige erinevamad muudest rahvastest. Muud rahvad on üksteise suhtes pigem geneetiliselt sarnased kui erinevad.

8. Diskussioon

Populatsioonigeneetika kinnitab geneetilisi erinevusi populatsiooni tasandil. See tuleneb näiteks nende rahvaste ajaloost, geograafilisest asukohast ja läbikäimisest naaberrahvastega. Antud peatükis arutleb autor töö praktilises osas saadud tulemuste üle; mida võib järeldada uuritud populatsioonide vahelistest seostest ja üleesindatud geneetilistest variantidest.

Ekstreemsete variantide uuringus avastati, et aafriklased ja ida-asiaadid on teiste võrdluspopulatsioonidega võrreldes omapärasemad. Neil on palju variante, mida teistes populatsioonides esineb väiksema sagedusega. Selline omapära on tingitud populatsioonide isolatsioonist teiste uurimispopulatsioonide suhtes. Isolatsioon tähendab, et suuremal osal rahval puudub kokkupuude teistest populatsioonidest pärit (geneetiliselt erinevama) inimestega. Ka soomlaseid käsitletakse populatsioonigeneetikas isoleeritud populatsioonina. Ühise andmebaasi uuringus (ExAC, 1000G, NLGo, UK10K, Eesti) märgati, et soomlased on samuti teistest populatsioonidest erinevamad. See tuleb soomlaste geneetilisest isolatsioonist, nimelt on nad ajaloo vältel geograafiliselt ja demograafiliselt olnud eraldatud ülejäänud maailmast [46]. Geograafiliselt lähedasemate populatsioonide korral näeme variantide ühtlasemat jaotumist, mis tuleb pidevamast ajaloolisest läbikäimisest [47].

1000G rahva ja kõikide andmebaaside ühisest peakomponendi analüüsist nähti samuti, et aafriklased on geneetiliselt kõige erinevamad ülejäänud populatsioonidest. Teised on omavahel pigem sarnased. Selline eristus on arvatavasti tingitud Aafrika elanike sisemisest geneetilisest mitmekesisusest. Inimkonna algkodu on Aafrikas. Väljarännanud populatsioonide geneetiline mitmekesisus on väiksem, kuna väljarännanud populatsioonid on endaga kaasa viinud vaid osa kogu sealsest variatsioonist [48]. Lisaks sellele tuleb aafriklaste erinevus muude populatsioonidega võrreldes sellest, et uuringus on vähe andmeid Aafrikale lähedal olevatest rahvastest. Sama kehtib ka Ida-Aasia põhiste rahvaste kohta – uuring näitab, et ida-asiaadid on geneetiliselt erinevad muudest populatsioonidest, sest käsitletud andmetes on vähe Ida-Aasia piirkonnas elavaid rahvaid.

Farmakogeneetilised variandid, mida esineb eestlastel rohkem või vähem kui teistel rahvastel, on igakülgsete ravimimetabolismi suunavate mõjudega. Mõned variandid tõhustavad kandjate ravimi metabolismi, teised aga vähendavad selle efekti. Eesti rahvastik on oma geneetiliselt

profiililt sarnased geograafiliselt lähedaste populatsioonidega nagu lätlased, leedukad, venelased [49].

Kokkuvõttes on käesoleva bakalaureuse töö käigus avastatud üksikuid huvitavaid seoseid erinevate maailma populatsioonide vahel. Uuringu teostamisel avastati, et Soomest, Aafrikast ja Ida-Aasia maadest pärit geneetilised andmed erinevad teiste võrreldavate populatsioonide geneetiliste sagedustega. Antud lõputöö käigus ei leitud, et eestlaste populatsioonis oleks ekstremaalsete farmakogeneetiliste variantide hulk ja sagedused märkimisväärselt erinevad lähedaste populatsioonide variantidest.

9. Kokkuvõte

Antud bakalaureusetöö eesmärk oli võrrelda farmakogeneetiliste variantide sagedusandmeid erinevates populatsioonides. Sooviti tuvastada eestlaste seas ekstremaalse levimusega variante. Varasem teadustöö on kinnitanud populatsioonides haiguste või ravimitarbimisega seotud variantide olemasolu. Selgitati geneetika teoreetilist tausta ja statistilisi meetodeid, kasutades avalikult kättesaadavaid geneetikaalaseid katalooge ja ressursse.

Uuringu käigus kirjeldati eestlaste seas ekstremaalse sagedusega geneetilisi variante. Nendel unikaalsetel variantidel on kõigil ka mingi kliiniline seos ravimitega. Kirjeldati viie variandi seoseid ravimitega nagu morfiin, tuberkuloosi ravim, metotreksaat ja statiinid.

Uuringu praktilises osas leiti andmete põhjal, et populatsioonid, mis asuvad geograafiliselt üksteisele lähemal, on geneetiliselt sarnasemad. Erandid olid aafriklased ja ida-asiaadid, mis eristusid nii üksteisest kui ka muudest populatsioonidest. See tuleneb ilmselt sellest, et need piirkonnad on teiste uuringus kasutatud populatsioonidega võrreldes geograafiliselt kaugemad ja teisi neile lähedasi populatsioone kasutati uuringus vähe. Leiti ka, et VIP ja mitte-VIP ning kodeerivate ja mittekodeerivate variandite sagedustes ei esinenud statistiliselt olulisi erinevusi.

10. Viidatud kirjandus

[1] A. von Bubnoff, „Next-generation sequencing: the race is on.," *Cell,* kd. 132, pp. 721-723, 7 March 2008.

[2] N. Wade, „NY Times," 4 9 2007. [Võrgumaterjal]. Available: http://www.nytimes.com/2007/09/04/science/04vent.html.

[3] „Understanding Human Genetic Variation," National Institutes of Health (US), 2007.

[4] R. L. Minster, N. L. Hawley ja C.-T. Su, „A thrifty variant in CREBRF strongly influences body mass index in Samoans," *Nature Genetics,* pp. 1049-1054, 1 September 2016.

[5] R. H. Villamil, M. Leon ja K. L. Macias, „A combined linkage and association strategy identifies a variant near the GSTP1 gene associated with BMI in the Mexican population.," *Journal of Human Genetics,* pp. 413-418, 1 March 2017.

[6] H. Cavanagh ja K. M. Rogers, „The role of BRCA1 and BRCA2 mutations in prostate, pancreatic and stomach cancers," *Hereditary Cancer in Clinical Practice,* p. 16, 1 August 2015.

[7] M. Qrafli, Y. Amar ja J. Bourkadi, „The CYP7A1 gene rs3808607 variant is associated with susceptibility of tuberculosis in Moroccan population.," *The Pan Americal medical journal,* p. 18, 1 March 2014.

[8] A. Heinaru, Geneetika, Tartu: Tartu Ülikool, 2012.

[9] L. Liu, „Comparison of Next-Generation Sequencing Systems," *Journal of Biomedicine and Biotechnology,* p. 11, 2012.

[10] M. A. Quail, „A tale of three next generation sequencingplatforms: comparison of Ion Torrent, PacificBiosciences and Illumina MiSeq sequencers," *BMC Genomics,* kd. 13, p. 341, 2012.

[11] A. Smertina, „Inimgenoomi ühenukleotiidiliste variatsioonide annotatsioon –ülevaade põhimõtetest ning teise põlvkonna sekveneerimise võimalike artefaktsete SNVdeannoteerimine," Tartu Ülikool, Tartu, 2016.

[12] C. G. v. El, „Whole-genome sequencing in health care," *European Journal of Human Genetics,* nr 21, pp. 580-584, 2013.

[13] M. Baker, „Biorepositories: Building better biobanks," *Nature,* kd. 486, pp. 141-146, 2012.

[14] „VCF variant call format," 2014. [Võrgumaterjal]. Available: http://www.internationalgenome.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40/.

[15] L. Leitsalu, T. Haller, T. Esko ja A. Metspalu, „Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu," *International Journal of Epidemiology,* pp. 1137-1147, 2015.

[16] K. Grand, Funktsioonikaoga mutatsioonide analüüs 2300 inimese genoomi ja terviseandmete põhjal, Tartu: Tartu Ülikool, 2016.

[17] N. Soranzo, „Nature," *526,* pp. 82-90, 1 10 2015.

[18] L. Francioli, „Nature Genetics," *46,* pp. 818-825, 2014.

[19] H. Lippmaa ja L. Trapido, Pärilikkusmeditsiin, Tallinn: Medicina kirjastus, 2010.

[20] C. E. Kelly, „Incorporation of Pharmacogenomics into Routine Clinical Practice: the Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline Development Process," *Current Drug Metabolism,* pp. 209-217, 1 Feb 2014.

[21] S. K. Balani, „Strategy of Utilizing In Vitro and In Vivo ADME Tools for Lead,“ *Current Topics in Medicinal Chemistry,* kd. 5, pp. 1033-1038, 2005.

[22] M. E. McDonagh, „From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource,“ *Biomarkers in Medicine,* kd. 5, nr 6, pp. 795-806, 11 December 2011.

[23] H. Chial, „DNA Sequencing Technologies Key to the Human Genome Project,“ *Nature Education,* p. 219, 2008.

[24] R. M. Durbin, „A map of human genome variation from population-scale sequencing,“ *Nature,* kd. 467, pp. 1061-1073, 28 10 2010.

[25] M. Lek, „Nature,“ *Analysis of protein-coding genetic variation in 60,706 humans,* pp. 25-291, 18 8 2016.

[26] HIM/NIGMS, „Dosign Guidelines - CPIC,“ 30 September 2014. [Võrgumaterjal]. Available: https://www.pharmgkb.org/view/dosing-guidelines.do?source=CPIC#. [Kasutatud 10 April 2017].

[27] P. Flicek, „About the Ensembl Project,“ 1 December 2016. [Võrgumaterjal]. Available: http://www.ensembl.org/info/about/index.html. [Kasutatud 1 March 2017].

[28] P. Flicek, „Ensembl 2012,“ *Nucleic Acids Research,* kd. 40, pp. 85-90, 15 November 2011.

[29] E. Sayers, „Information, Database resources of the National Centerfor Biotechnology,“ *Nucleic Acids Research,* kd. 39, pp. 38-51, 4 November 2011.

[30] L. I. Smith, „A tutorial on Principal Components Analysis,“ 2002.

[31] S. Suyash, „Statistical Methods for studying Genetic Variation in Populations,“ Carnegie Mellon University, Pittsburgh, 2012.

[32] S. Wright, Evolution and the Genetics of Populations. Vol. 2, The Theory of Gene Frequencies, Chicago: University of Chicago, 1969.

[33] D. McDonald, „Worked example of calculating F-statistics from genotypic data:,“ 1 September 2013. [Võrgumaterjal]. Available: http://www.uwyo.edu/dbmcd/popecol/maylects/fst.html. [Kasutatud 22 March 2017].

[34] M. Abramowitz ja I. A. Stegun, „Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables,“ Washington D.C, Tenth Printing, 1965, p. 470.

[35] R Foundation, „R Studio,“ 5 January 2015. [Võrgumaterjal]. Available: https://www.rstudio.com.

[36] EMBL-EBI, „Sample 1000 Genomes,“ 1 January 2008. [Võrgumaterjal]. Available: http://www.internationalgenome.org/category/sample/. [Kasutatud 11 April 2017].

[37] Broad Institute, „ExAC Browser Frequently Asked Questions,“ 20 October 2014. [Võrgumaterjal]. Available: http://exac.broadinstitute.org/faq. [Kasutatud 11 April 2017].

[38] Y. Xiang, L. Ma, W. Wu, W. Liu, X. Zhu, Q. Wang, J. Ma, M. Cao, Q. Wang, X. Yao, L. Yang, A. Wubuli, C. Merle, P. Milligan, Y. Mao, J. Gu ja X. Xin, „The incidence of liver injury in Uyghur patients treated for TB in Xinjiang Uyghur autonomous region, China, and its association with hepatic enzyme polymorphisms nat2, cyp2e1, gstm1 and gstt1,“ *PLoS ONE,* kd. 9, nr 1, p. 1, 1 Jan 2014.

[39] T. Fukuda, V. Chidambaran, T. Mizuno, V. Olbrecht ja A. Vinks, „OCT1 genetic variants influence the pharmacokinetics of morphine in children.,“ *Pharmacogenomics,* pp. 1141-1151, 1 July 2013.

[40] M. Ferrari, L. Guasti, A. Maresca, M. Mirabile, S. Contini, A. Grandi, F. Marino ja M. Cosentino, „Association between statin-induced creatine kinase elevation and genetic polymorphisms in SLCO1B1, ABCB1 and ABCG2.," *European journal of clinical pharmacology,* pp. 539-547, 2014.

[41] L. Hughes, T. Beasley, H. Tiwari, S. Morgan, J. Baggott, K. Saaq ja J. McNicholl, „Racial or ethnic differences in allele frequencies of single-nucleotide polymorphisms in the methylenetetrahydrofolate reductase gene and their influence on response to methotrexate in rheumatoid arthritis.," *Annals of the rheumatic diseases,* pp. 1213-1218, 2006.

[42] R. Ho, „Effect of drug transporter genotypes on pravastatindisposition in European- and African-American participants," *Pharmacogenet Genomics,* pp. 647-656, 2007.

[43] M. Pasanen, T. Miettinen, H. Gylling, P. Neuvonen ja M. Niemi, „Polymorphism of the hepatic influx transporter organic anion transporting polypeptide 1B1 is associated with increased cholesterol synthesis rate," *Pharmacogenetics and Genomics,* pp. 921-926, 1 October 2008.

[44] J. Chung, J. Cho, K. Yu, J. Kim, K. Lim, D. Sohn, S. Shin ja I. Jang, „Pharmacokinetic and pharmacodynamic interaction of lorazepam and valproic acid in relation to UGT2B7 genetic polymorphism in healthy subjects," *Clinical pharmacology and therapeutics,* kd. 83, nr 4, pp. 595-600, 1 April 2008.

[45] J. Johnson, I. Zineh, B. Puckett, S. McGorray, H. Yarandi ja D. Pauly, „Beta 1-adrenergic receptor polymorphisms and antihypertensive response to metoprolol.," *Clinical pharamacology and theurapeutics,* pp. 44-53, 1 July 2003.

[46] J. U. Palo, „Genetic Markers and Population History: Finland Revisited," *European Journal of Human Genetics,* pp. 1336-1346, 2009.

[47] Public Library of Science, „ScienceDaily," 7 June 2009. [Võrgumaterjal]. Available: www.sciencedaily.com/releases/2009/06/090605091157.htm. [Kasutatud 29 March 2017].

[48] S. Tishkoff, F. Francoise, F. A. Reed, F. R. Friedlaender ja C. Ehret, „The Genetic Structure and History of Africans and African Americans," *Science,* pp. 1035-1044, 2009.

[49] T. Esko, „Novel applications of SNP array data in the analysis of the genetic structure of Europeans and in genetic association studies," University of Tartu Press, Tartu, 2012.

[50] M. Scholz, „Approaches to analyse and interpret," Max-Planck-Institut f¨ur Molekulare Pflanzenphysiologie, Potsdam, 2006.

[51] The R Foundation, „About R Studio," 10 September 2002. [Võrgumaterjal]. Available: https://www.r-project.org/about.html.

[52] R Foundation, „R Studio," [Võrgumaterjal]. Available: https://www.rstudio.com/.

[53] Herstein, Topics In Algebra, New York: John Wiley & Sons, 1964.

[54] J. Pazik, M. Oldak, M. Dabrowski, Z. Lewandowski ja E. Sitarek, „Association of UDP-glucuronosyltransferase 1A9 (UGT1A9) gene polymorphism with kidney allograft function," *Annals of transplantation,* pp. 69-73, 30 October 2011.

11. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, **Kevin Kanarbik**,
(autori nimi)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
, **Farmakogeneetilised variandid täisgenoomsetes andmetes**

   mille juhendaja on **Tõnis Tasa**,
   (juhendaja nimi)

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **05.05.2017**